5-1-2019

Twitter Bot Detection in the Context of the 2018 US Senate Elections

Wes Kendrick
Dartmouth College

Twitter Bot Detection in the Context of the 2018 US Senate Elections

Wes Kendrick

Senior Thesis in Computer Science

Dartmouth Computer Science Technical Report TR2019-865

Advisors

Professor V.S. Subrahmanian

Professor Benjamin Valentino

May 2019

Abstract

A growing percentage of public political communication takes place on social media sites such as Twitter, and not all of it is posted by humans. If citizens are to have the final say online, we must be able to detect and weed out bot accounts. The objective of this thesis is threefold: 1) expand the pool of Twitter election data available for analysis, 2) evaluate the bot detection performance of humans on a ground-truth dataset, and 3) learn what features humans associate with accounts that they believe to be bots. In this thesis, we build a large database of over 120 million tweets from over 900,000 Twitter accounts that tweeted about political candidates running for US Senate during the 2018 American Midterm Elections. Tweet-level data were collected in real-time during the two-month period surrounding the elections; account-level data were collected retrospectively in the months following the elections. Using this original dataset, we design and launch a bot detection study using a novel combination of Amazon SageMaker and Qualtrics. For ground truth, we include 39 known bot accounts from a separate 2015 Bot Challenge Dataset (BCD 2015) in the study sample. Of the 39 known bots from BCD 2015, only 11 accounts (28.2%) were accurately identified as bots with a two-thirds or unanimous annotator vote; just 5 accounts (12.8%) were unanimously accurately identified as bots, highlighting the difficulty of building accurate training sets for bot detection. Looking at the study results for the Senate dataset accounts, we observe that accounts which 1) post frequently and 2) retweet frequently were more likely to be labeled as bots. The Senate dataset and the associated study results offer significant opportunities for further analysis and research.

Contents

Chapter 1

Introduction

1.1 Motivation

Free and fair elections form the foundation of a well-functioning democracy. In today's networked age, political messages travel rapidly around the world with just a few clicks, disrupting traditional top-down mass media. Now, anyone, or anything, can reach millions of users.

In the aftermath of the 2016 election, computer and social scientists alike both sought to understand precisely what happened during the election, and to what extent outside actors managed to influence the conversation online. Rather than add to politicized existing literature, we decided to pursue original research on a new set of elections, the 2018 Midterms, as they happened in real-time. We picked the 35 elections for US Senate in particular because they represented a specific, manageable subset of races to gather data on. Then, using our dataset, we crowdsource human judgements regarding whether or not Twitter accounts are controlled by humans or by bots.

An ancillary motivation for this work is to improve tooling and resources for analysis. The state of the art for social scientists up to this point has generally been to upload CSVs to Amazon Mechanical Turk or comparable platforms, but manual CSV uploads are fairly limited, especially in terms of the amount of data that can realistically fit on a single row. From the computer science side, researchers com-

monly use human annotators (coders, in political science terms) to build training datasets for machine learning problems, but those training datasets are typically of fairly simple and concrete things – detecting dogs versus cats, as a simple example. Bot detection is a far more difficult challenge, especially because labeling is prone to bias. With bot detection, our novel contribution has been to work to combine social science methodology (collecting demographic data, carefully designing an intervention to minimize bias) with more advanced tooling used to develop machine learning datasets. Bot detection is especially difficult because we almost always lack ground truth. A final theme and motivation throughout this work is constant care to avoid making unwarranted assumptions about what is or is not a bot account. On the Senate dataset, we lack ground-truth, so we must not get carried away with grand claims. This dataset is not a random sample of Twitter generally, and thus the results here may not generalize beyond this particular selection of Twitter accounts. Caveats aside, we can say a few interesting things about what people think bots are, because accounts with certain attributes were indeed more likely to be perceived to be bots. Of course, we do not know for certain if they actually are bots; we only know that a sample of humans thought they were bots.

1.2 Related Work

We began by studying the literature, looking for prior examples of bot detection, large-scale data collection, and large-scale crowdsourcing. First, in terms of bot detection, researchers clustered around Indiana University, including E. Ferrara, A. Flammini, and F. Menczer, have built an interesting bot detection service (`https://botometer.iuni.iu.edu/`) as part of their broader Observatory on Social Media (OSoMe) research project `https://osome.iuni.iu.edu/`. For more information on their bot detection research, refer to [1] and the more-recent [2]. E. Ferrara, A. Flammini, and F. Menczer were also involved with the 2015 Twitter bot challenge [3]. We use the known bot accounts from the Twitter bot challenge dataset (referred to in this paper as BCD 2015) as ground truth in this thesis. Refer to section 2.2.3

for more information about how we use BCD 2015.

The same group of researchers, Ferrara et al. (2016) offer a review of recent literature discussing social bots [4]. In reviewing the history of known bot campaigns, they refer back to the 2010 midterm elections, and they discuss how bots were employed to boost some candidates and harm others. They propose a taxonomy of bot detection systems: (none of these are mutually exclusive):

1. social network information
2. human intelligence and crowdsourcing
3. machine learning

Twitter rate limits make it relatively difficult to get graph network information from Twitter's API (follower and followee lists in particular are locked down, for good reasons), which makes it difficult to get the necessary data to run a classifier using network-based features. "Human Intelligence and Crowdsourcing" is the second general approach on their taxonomy, and it is the primary approach used here in this paper. Before building a model, the third item on their list, it's important to understand and have good, accurate training data – supervised machine learning models are only as good as the quality of the training data which is input. We will attempt ML in future work. Finally, the paper observes that the bot landscape is constantly changing [4].

Because Twitter data collection is a relatively low-barrier data source, many groups, private and public, have done large-scale data collection. From the published literature, T. Finin et al. (2010) builds a nice system to crowdsource named entity recognition [5]. In their paper, they weigh the trade-offs between Amazon Mechanical Turk and CrowdFlower (now rebranded as FigureEight). In 2010, CrowdFlower was likely the best tool available; now Amazon SageMaker offers a compelling alternative choice, and SageMaker is the tool used in the present work. Finin et al. discusses the use of a 'gold standard dataset', which they use to monitor annotator performance and enforce standards. In their case, because named entity recognition is a right-or-wrong task, you're either correct or incorrect, and an outside observer can indeed

verify that a particular result is correct, the golden dataset approach makes sense – annotators that perform below par on a known task can be removed from the study. Whereas in this study, because we don't fundamentally know which Senate accounts are or are not bots, we avoid removing any annotators from our labeling pool. We do, however, track and associate all responses with demographic data, so subsequent cohort analysis will be possible.

Chu et al. (2010) discusses the design and implementation a multi-class classifier capable of detecting human, bot, and cyborg accounts. The work broadens the traditional bot-human binary classification, pushing back on an assumed binary that an account is either a bot or a human [6]. Chu's work serves as an inspiration for why, in the present work, we require that annotators who indicate that an account is a bot also provide a bot subtype, if they end up analyzing an account which they believe to truly be a bot. Interesting future work would be to include 'cyborg' as an option for annotators.

Cohen and Ruths (2013) ask the question: do political orientation classifiers perform as well as their authors claim they do, especially on less-polar Twitter accounts [7]? The authors find that many political orientation classifiers do not generalize beyond their narrow sample of tweets or accounts that they were trained on; a cautionary tale with regard to overly-optimistic estimates of model quality. In their work, they test several political orientation classifiers on two sets of accounts: 1) politicians and 2) 'normal' users who rarely discuss politics, and finds that inference accuracy on normal users is far below what some researchers have claimed in their published results [7]. Skewed training data degrades classifier performance and limits generalizability. Thus, the focus here has been on building a system to reliably collect good data, first, rather than rushing to build a model on data that we do not fully understand. The relevance of this for this paper is that, while it is tempting to build a bot detection model straight away, there first are many important questions that we must first ask about ourselves and our data, before we can come close to generalizing.

1.3 Research Questions

The following are guiding questions we used to structure our research:

1. What did the 2018 Senate Elections look like on Twitter?
2. How good are humans at detecting bots?
    - What are bots?
    - How do we know if an account is a bot?
    - How will we show people Twitter accounts?
3. What makes people think that certain accounts are controlled by bots, but not others?

We will return to these questions throughout this paper.

1.4 2018 Elections for US Senate

The 2018 American Midterm Elections took place on November 6, 2018. There were a total of 35 Senate races: 33 regular races and 2 special elections.

The MIT Election Lab hosts a great repository of datasets with election data for the curious; you can find election statistics, including 2018 Midterm data, on the MIT Election Lab Website [8].

1.5 State of the Art: Building Datasets for Machine Learning

A number of tools and services exist with the purpose of building training datasets.

Within large tech companies, in-house labeling pipelines are commonplace and are integrated into complex internal workflows.

Amazon Mechanical Turk remains the leading human-tasks platform, although there are many new competitors. The platform was not built from the ground up

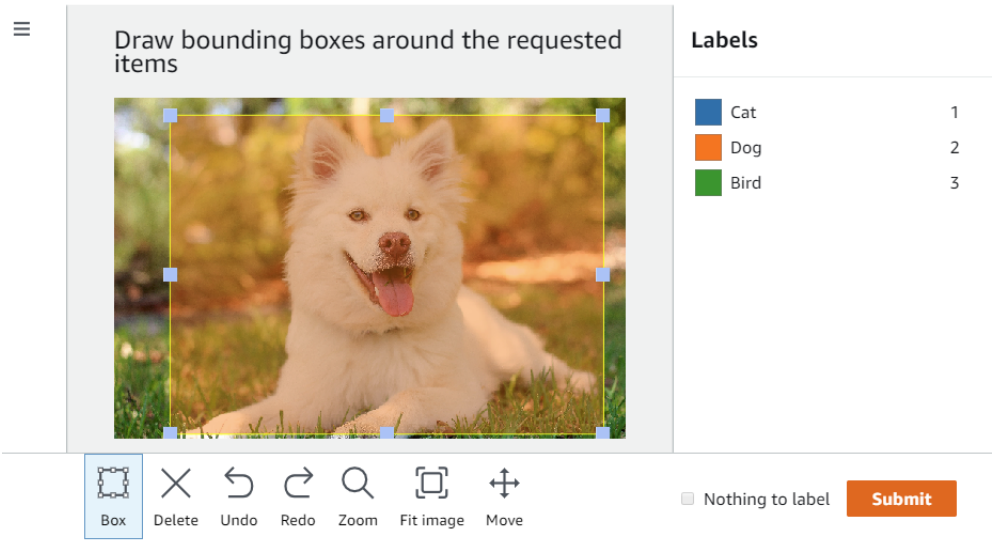


Figure 1.1 MTurk Standard Labeling Interface for Object Detection [11]

to serve as a ML labeling tool per se, although it also functions quite well for ML labeling. Other, newer alternatives are designed from the ground-up for ML labeling, including FigureEight, LabelBox, and many others.

Prodigy is a relatively new entrant [9]. It uses active learning to reduce the number of annotations necessary to train a supervised learning classifer. It is strong as a data exploration tool for data scientists. Because it does not (yet) support multiple annotators per object out of the box, it is not as well suited for tasks that are prone to bias, such as this bot research.

SageMaker Ground Truth is a new service from Amazon Web Services which is useful for building datasets for machine learning [10]. You can run SageMaker labeling jobs on three types of workforces: 1) on Amazon Mechanical Turk, with a 3rd party vendor or 3) with your own team.

SageMaker Ground Truth is quite generalizable – you can add annotations to just about any type of data that you can save to Amazon S3. Several stock annotation templates are provided out-of-the box; for more complex annotation tasks, such as the present task, they provide the option to build-your-own labeling task using your own HTML+CSS+Javascript.

Bias in machine learning is a major concern; see, for example, Dressel's thesis

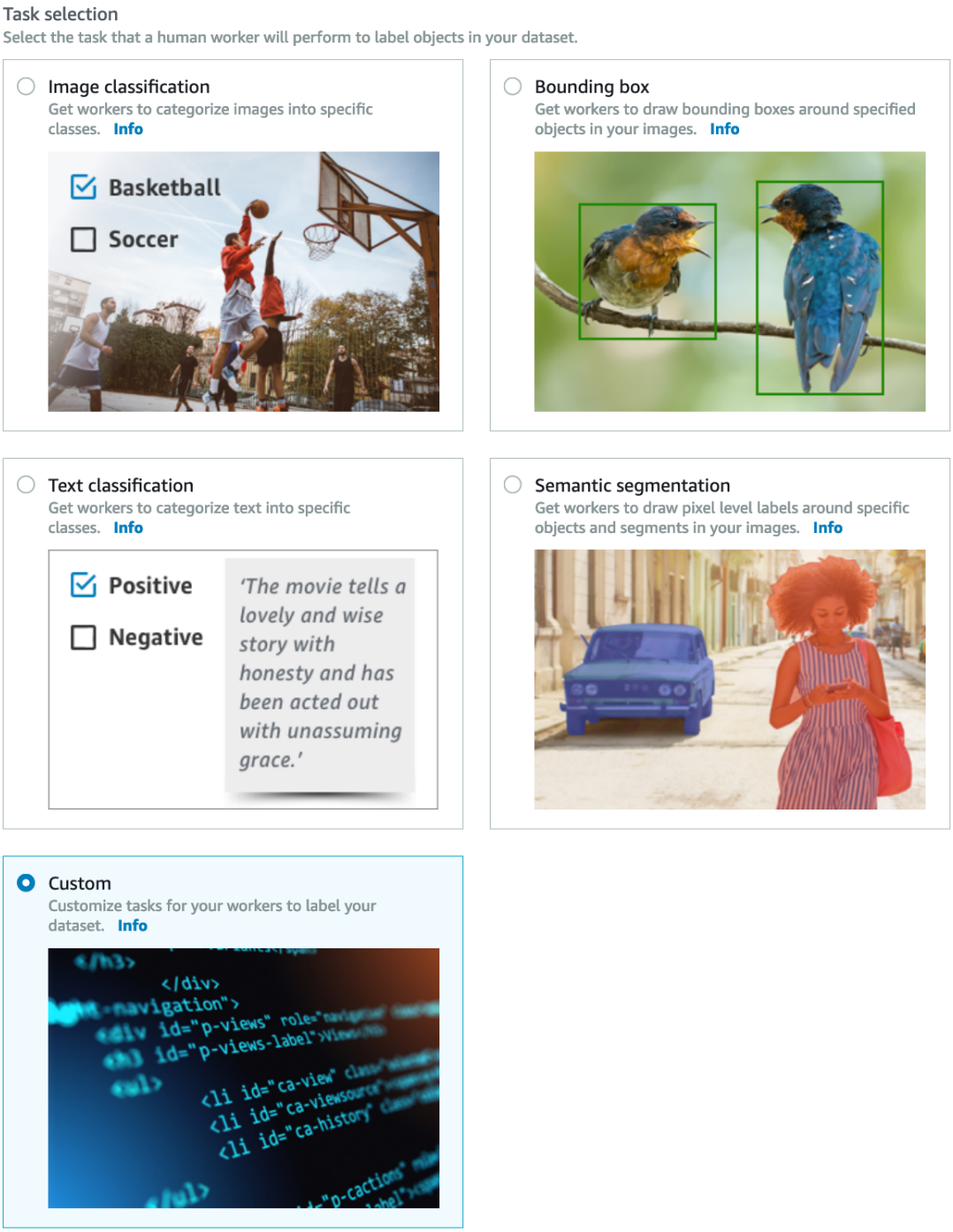


Figure 1.2 For this research, we implemented a custom bot detection task using AWS SageMaker

work analyzing racial bias in a recidivism algorithm [12]. Bias from the training set is perpetuated with the model, so catching the bias in the dataset from the very start is an important task.

1.6 Data Science Ethics

This research involves over a terabyte of text data, and it is important that we are careful stewards of it. For this research, we received all necessary approvals – we received IRB approval (Exempt – Category 2) and department approval. All data downloaded was via an authorized Twitter Developer Account at the free-tier public level. No scraping was involved – scraping is potentially a violation of Twitter's terms of service. All tweets were archived from public accounts. If we ever make any of the dataset public, we will only make the tweet IDs themselves public, in accordance with Twitter's Tweet Rehydration Policy [13]. Digital regulation is relatively immature, and we expect regulations to change and evolve to better protect and inform people over the coming years. We seek to comply with the spirit and letter of applicable regulations.

Chapter 2

Methods

2.1 Data Archive Service

In order to archive large amounts of data, we rented an Amazon Web Services (AWS) EC2 instance and a MySQL database on RDS for the duration of the study. The EC2 instance ran a Python script as a cron job. The script hit the Twitter API every fifteen minutes and downloaded tweets which matched the specified search queries up to the rate limits. Refer to Appendix A for a list of search queries.
At a slightly lower-level, in addition to everything mentioned above, we also use S3 for object backup. On the development machine (locally): MongoDB and Python pickles for object storage to disk. Saving data on a local disk makes a lot of sense, as it allows us to avoid making expensive and slow queries over the network repeatedly. Lawson discusses interesting work on Exascale computing which would be relevant for an even larger data collection effort [14].

2.2 Datasets

There are three datasets in this research, two of which we build:

1. Senate Real-Time Dataset

2. Senate Retrospective Dataset

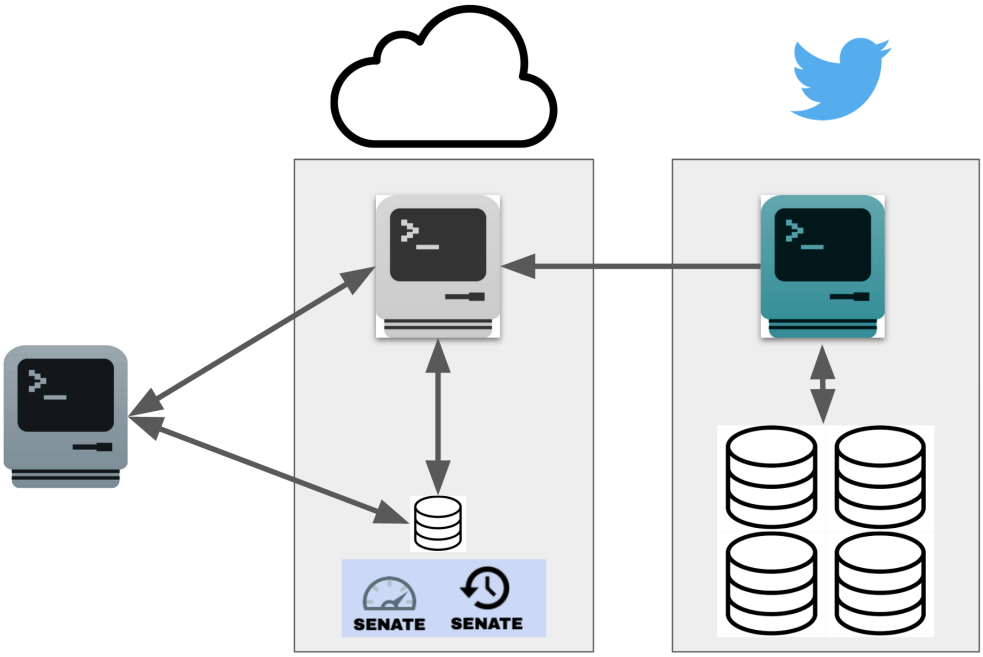


Figure 2.1 High-level service overview

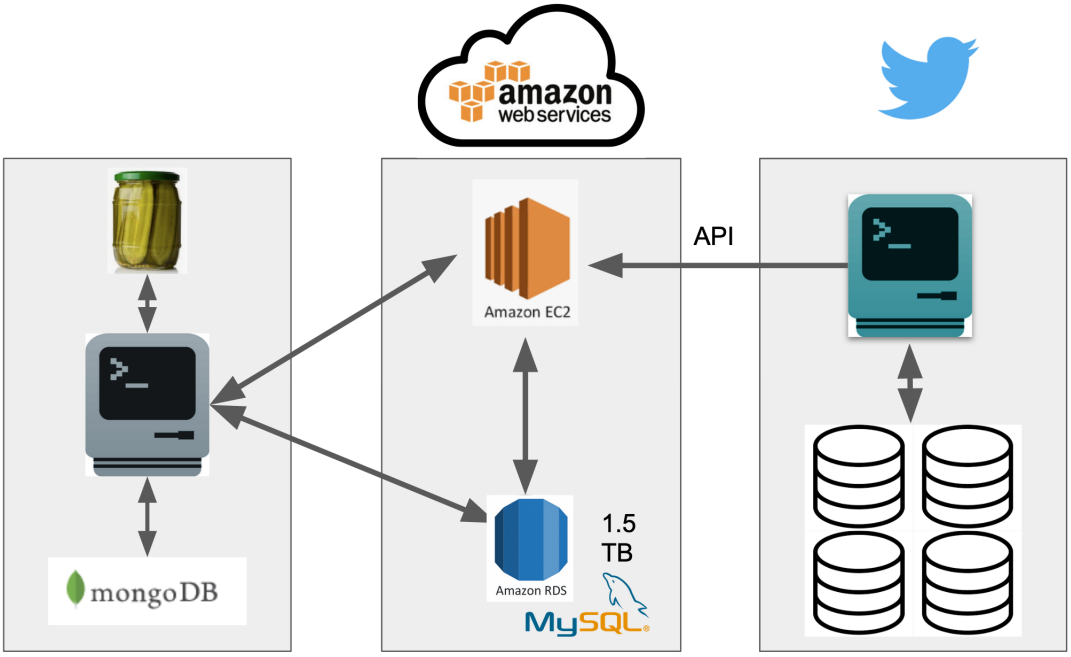
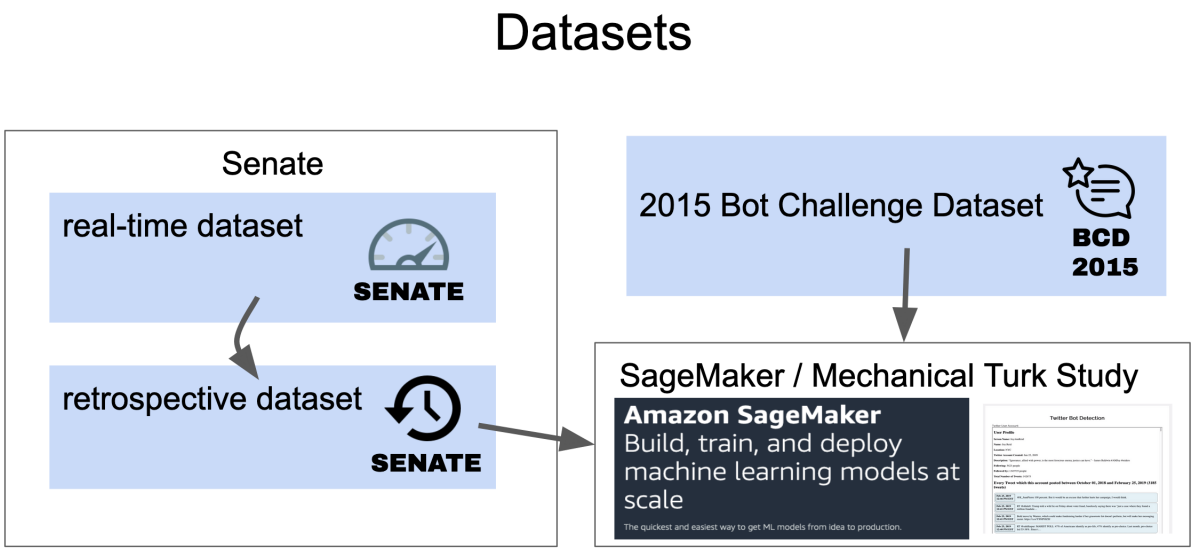


Figure 2.2 Detailed service overview

3. 2015 Bot Challenge Dataset (BCD 2015) [3]

The following subsections will explain in more detail the methodology underlying each of the three datasets.



2.2.1 Senate Real-Time Dataset

The only real-time Senate dataset of the study, this script ran as follows:

- Start Collection: October 14, 2018
- Stop Collection: December 10, 2018
- 5,426,083 tweets
- 947,099 unique twitter accounts

Important to note that this is first dataset was collected at the tweet level – it hits the Twitter search API endpoint with the queries listed in Appendix A, up to but not exceeding the free-tier dev limit. The data collection started as soon as we were ready to run the tweet collector live.

Search Type	Format	Example Search
Keyword Search	Firstname Lastname	Rick Scott
Hashtag Search	#FirstnameLastname	#RickScott
Tweets Posted By Senate Candidate Twitter Accounts	@username	@ScottforFlorida

Figure 2.3 Real-Time Dataset API Query Types

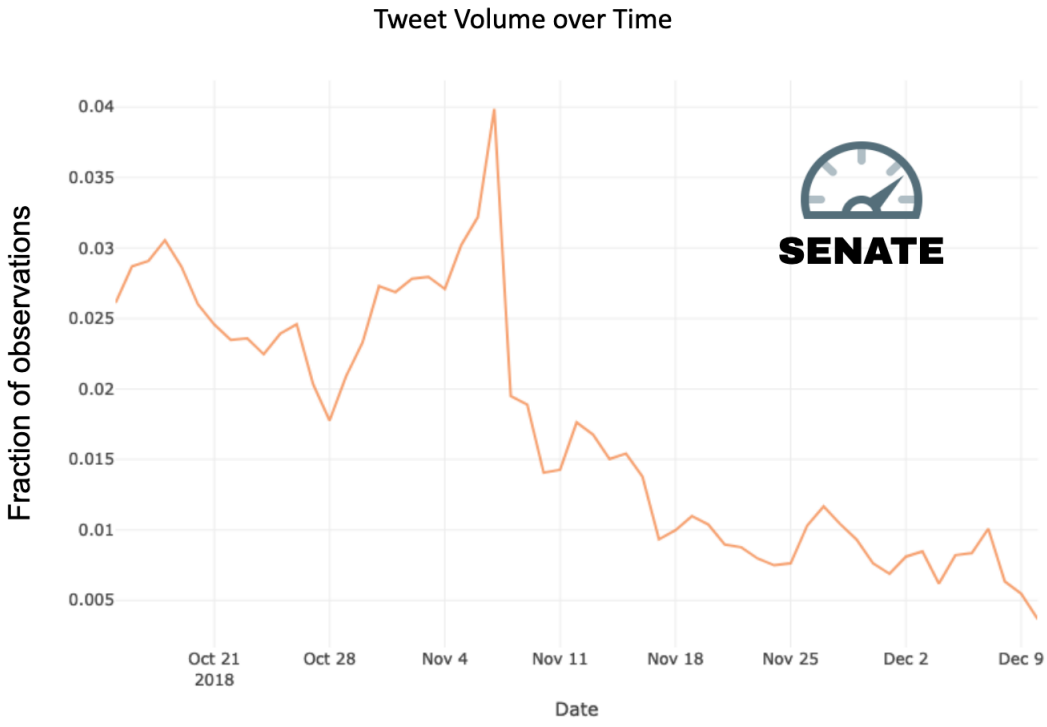
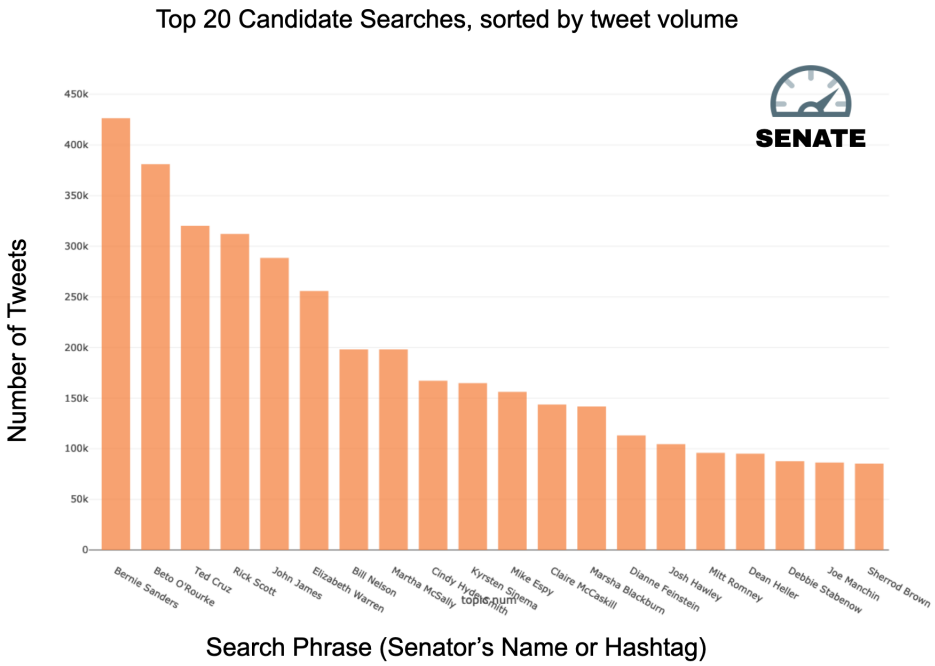
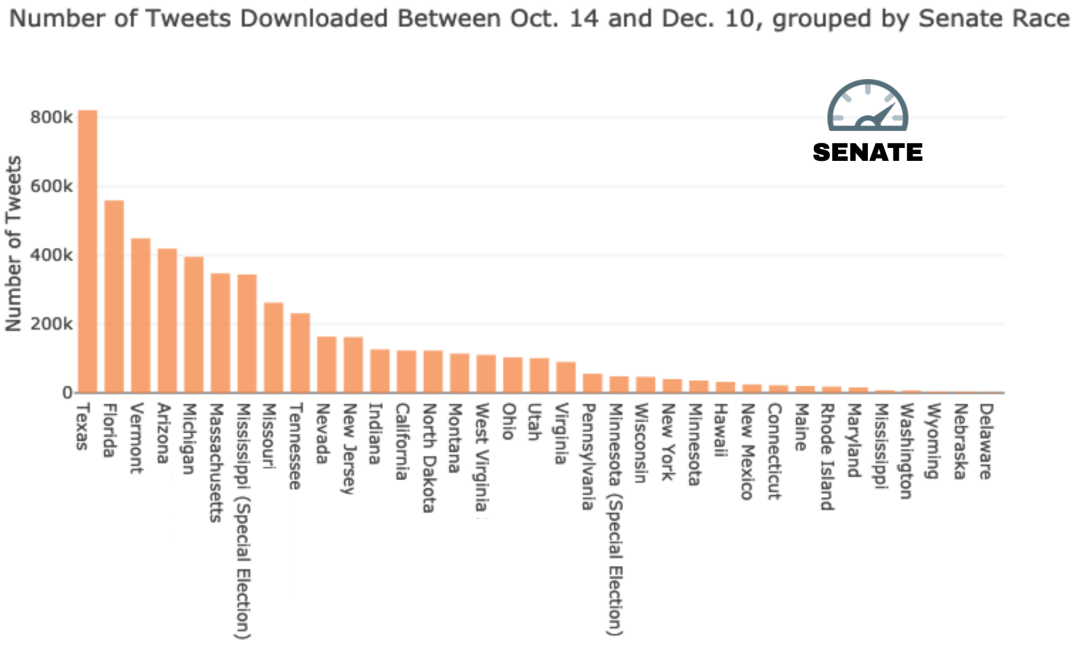


Figure 2.4 Tweet volume over time – large spike is from the Nov. 6 elections



Some candidates have national followings, for example Bernie Sanders, who also happened to be up for election for the state of Vermont.



Looking at volume of tweets by Senate race.

Word	Frequency	Word	Frequency	Word	Frequency
vote	640,392	senate	614,395	democrat	500,615
beto	440,884	trump	422,818	cruz	384,479
ted	364,438	bernie	359,462	sanders	336,411

Figure 2.5 Most common words in tweet text, sorted by volume. These results are intuitive and largely confirm our expectations; we are querying for candidate names and we logically see the most popular candidates represented here, along with common election-related words.

2.2.2 Senate Retrospective Dataset

Up to this point, the data downloaded has been at the tweet level. However, if we are to detect bot accounts, we need contiguous selections of tweets from single accounts in order to produce account profiles. Of the approximately 947,000 accounts captured in the real-time batch, we initially attempted to download a sample of tweets from every account referenced in the real-time database. However, that quickly proved infeasible due to rate limits, and would have taken more than a year to complete. Thus, we determined that the next-best thing would be to pull a random sample. As such, we pulled a random sample of 85,000 accounts from the list of roughly 947,000 accounts present in the Senate realtime database.
In order to build the retrospective dataset, we used the GET statuses/user_timeline endpoint. The endpoint works as follows: it allows one to query for tweets in reverse-chronological order starting from the current day and going back in time until either:

- since_id parameter, which we set to October 1, 2018, or
- The API maxes out and only displays the 3200 most recent tweets/account

(Whichever comes soonest)

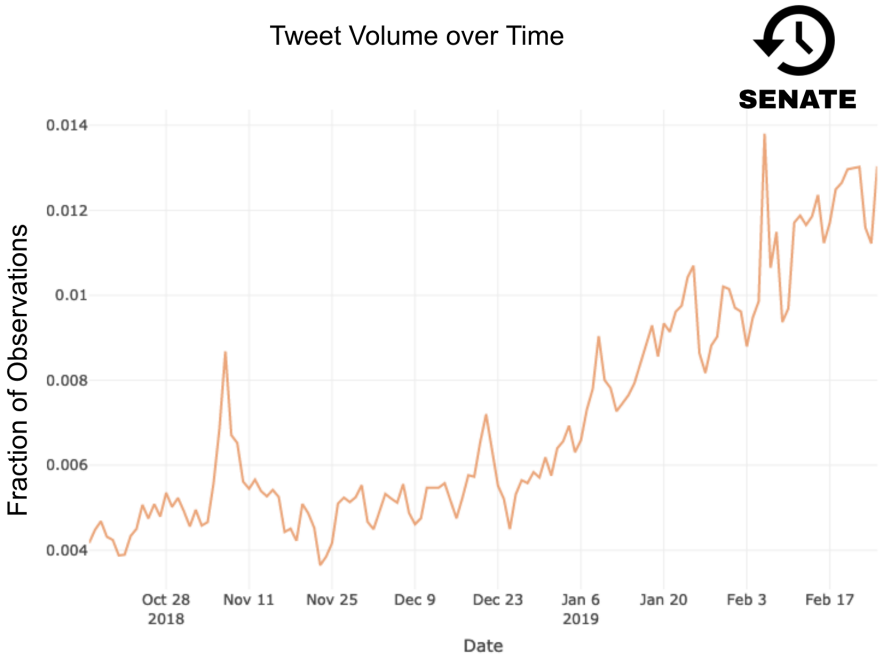


Figure 2.6 Post volume over time, represented as a timeseries. The volume of tweets collected over time skews towards later timestamps due to API limitations. Tweet volume is not increasing over time; volume of collected data clearly is. Many accounts have tweeted frequently enough that their old tweets are no longer accessible. Despite the skewed data collection, we can still notice two major spikes: 1) Election on Nov. 6, 2018 and 2) State of the Union on Feb. 5, 2019

The Senate retrospective job ran from February 26, 2019 - April 1, 2019, given the rate limit constraints. At job completion, 128,210,831 tweets from 85,000 accounts were archived, taking up over 800 gigabytes on disk. Because of the rate limits, the download took months.

A significant artifact from the data collection process is visible in the chart. Many of the accounts in the sample post frequently. If they post more than 3,200 times between February 2019 and October 2018, the API will max out and stop giving older tweets once we hit the cap. As a result, we have more tweet data in the datset for later months, and less data corresponding to earlier months. So while it seems like the volume of tweets over time is increasing, that is not the case. It just reflects the distribution of the timestamps of the data contained in the database.

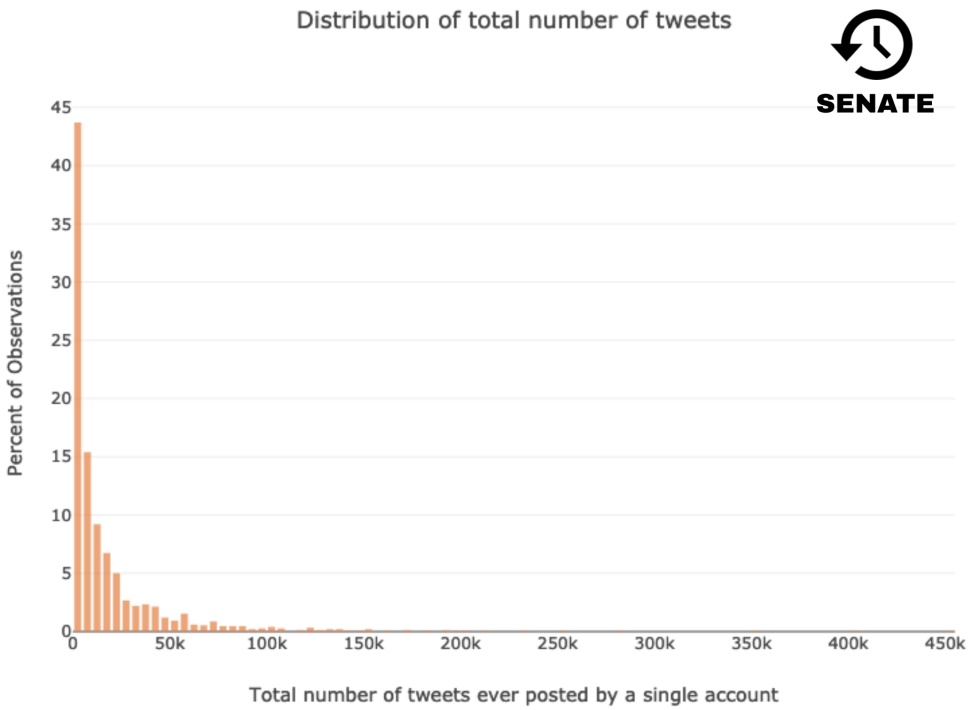


Figure 2.7 Distribution of all-time total number of tweets, by account.

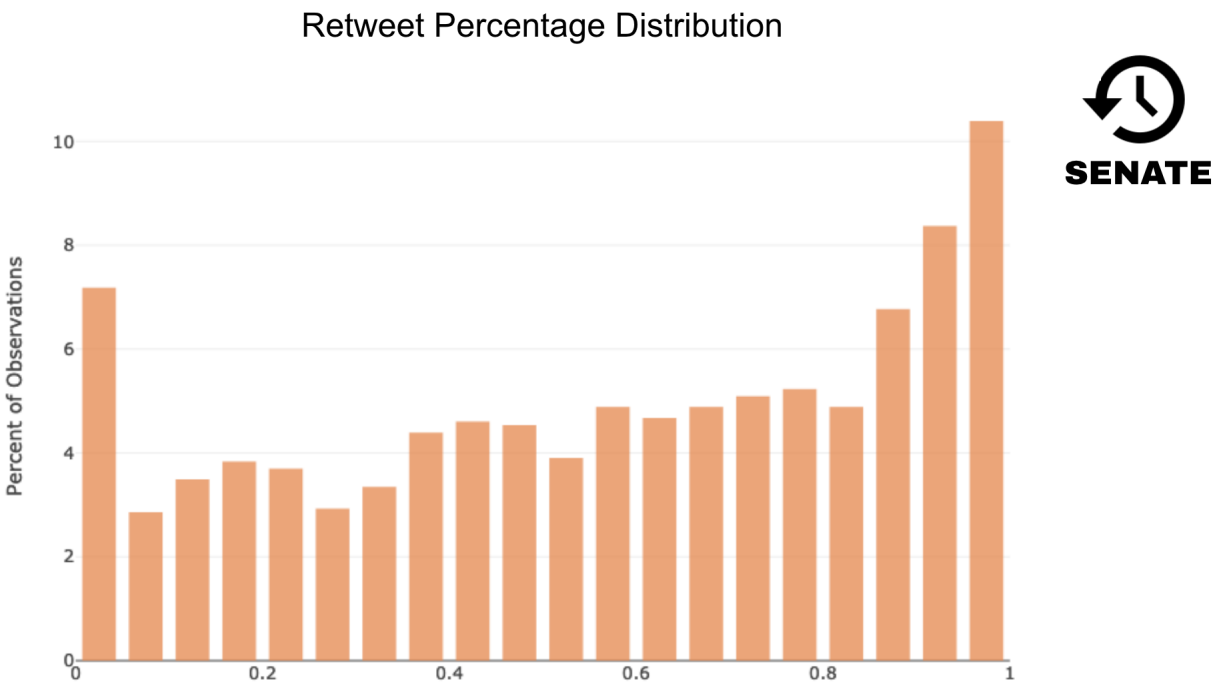


Figure 2.8 Note that some accounts in the sample have zero tweets (and thus by definition also zero retweets), they fall into the leftmost bar in the chart and have not been filtered out from the visualization.

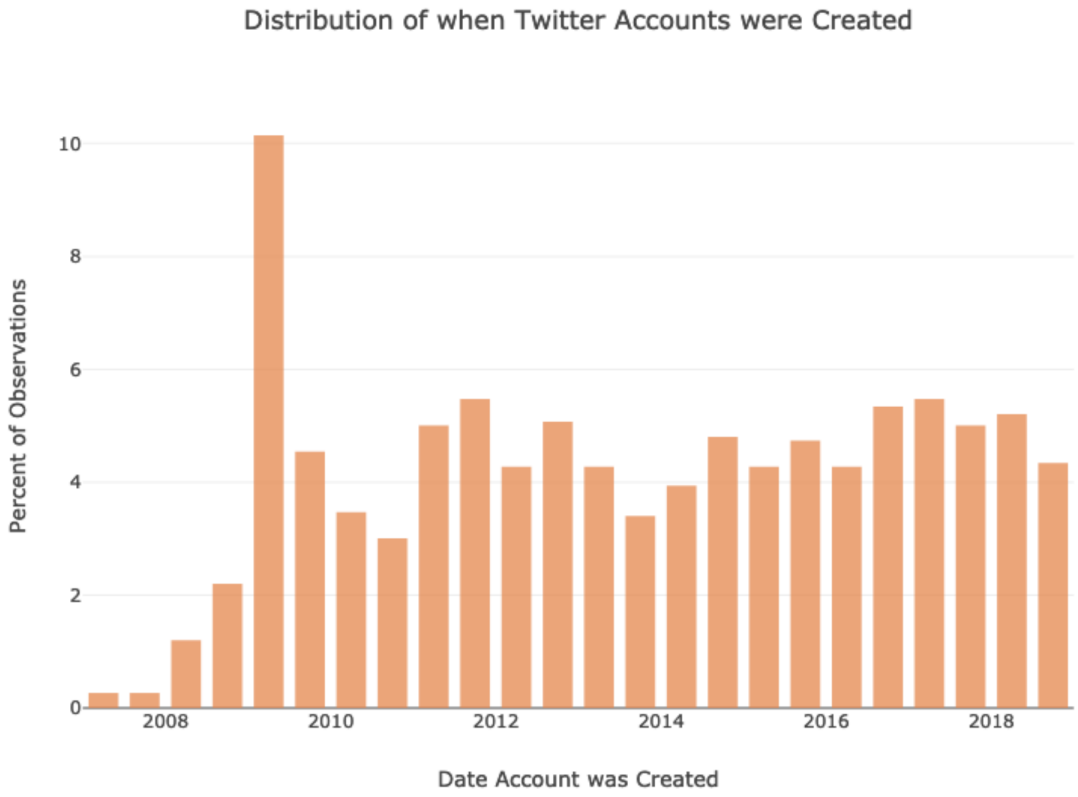


Figure 2.9 Distribution of when Twitter accounts were created is fairly uniform, except for an unexplained spike of accounts created around 2009

2.2.3 2015 Bot Challenge Dataset (BCD 2015)

The third dataset comes from Professor V.S. Subrahmanian's prior research. Several university and corporate teams competed to detect a set of pro-vaccination Twitter bots in the BCD 2015 dataset. Professor Subrahmanian led the team which won first place at the competition with a near-perfect score. His winning team used a semi-automated process, integrating inconsistency detection and behavioral modeling, text analysis, network analysis, and machine learning to detect bots [3]. General BCD 2015 Statistics:

- 7,038 accounts, of which 39 are known pro-vaccination bots
- 4,095,083 tweets
- Network data
- The only thing we know is that those 39 accounts are bots. It does not follow that the remaining accounts are not bots – they might also be bots.

2.3 Study Design

At a high-level, we do the following:

1. Generate HTML files of Twitter accounts, displaying profile information and a selection of tweets from data and metadata stored in our database. The Multi-page conditional form is based on the following template from W3Schools.com [15].
2. Build a demographic survey, which annotators are required to take before they can begin the annotation task.
3. Build a custom SageMaker labeling task, which shows annotators the HTML Twitter file in an iframe, and asks a series of questions via a conditional webform

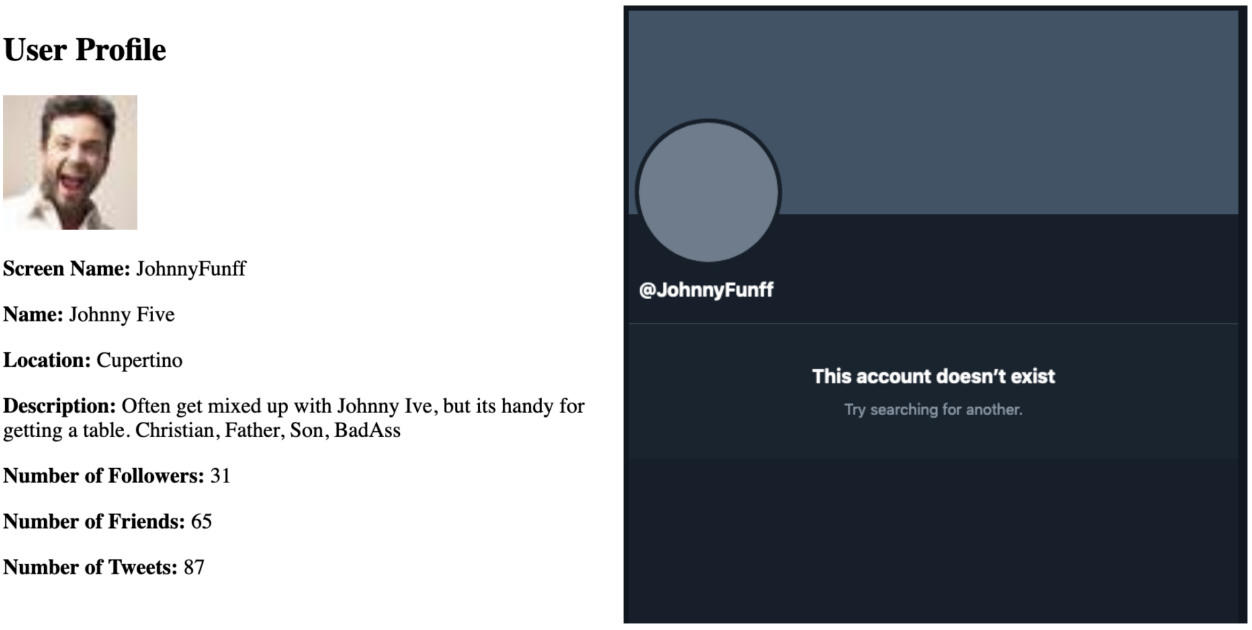
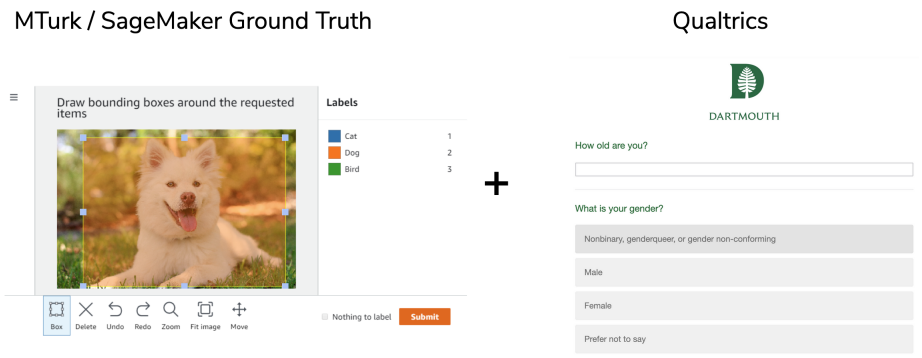


Figure 2.10 Left: generated known bot html file from BCD 2015. Right: screenshot of account as of May 2019, notice that it no longer exists. Interesting to note that the profile image shown on the left is still being hosted on Twitter's image servers, and is accessible even today if you know the image URL to query for. You might expect that Twitter would delete those assets to save space; we find that is (experimentally) not always the case.



2.3.1 Qualtrics Demographic Survey

We asked the following questions:

1. Study Consent ((yes), (no))
2. How old are you (numeric)
3. What is your gender? ((male), (female), (Nonbinary, genderqueer, or gender non-conforming), (Prefer not to say))
4. Generally speaking, do you think of yourself as a? ((Strong Republican), (Weak Republican), (Lean Republican), (Independent), (Lean Democrat), (Weak Democrat), (Strong Democrat), (Other free-text), (Prefer not to say))
5. What is the highest level of education that you are currently enrolled in or have obtained? ((High School), (Trade School), (Associate's Degree), (Undergraduate Degree), (Masters Degree), (Doctorate / PhD), (Prefer not to say))
6. Do you use Twitter? If so, how frequently? ((Never), (Monthly), (Weekly), (Daily), (Multiple times per day))
7. How many tweets have you posted within the last 30 days (including retweets) ((0 tweets), (Between 1 and 10 tweets), (More than 10 tweets))
8. How frequently do you read the news? ((Never), (Monthly), (Weekly), (Daily), (Multiple times per day))

9. How do you get your news? (click all that apply) (can select multiple) ((Print Newspaper), (Magazines), (TV), (Computer / Smartphone / Internet), (Other free-response), (Prefer not to say)
10. Would you say you follow what's going on in government and public affairs: ((Rarely), (Sometimes), (Most of the time), (Almost always), (Prefer not to say))
11. How would you define what a bot on Twitter is? (Free-text response)
12. What percentage of accounts on Twitter.com do you think are controlled by bots? (select percentage between 0-100)
13. Do you think bots on Twitter are mostly helpful or mostly dangerous?
14. How helpful/dangerous do you think bots on Twitter are? ((A little helpful), (Somewhat helpful), (Very helpful)) (conditional question based on previous answer)
15. How are bots on Twitter helpful/dangerous (Free response)
16. How confident are you in your ability to identify bot accounts on Twitter? ((Not at all confident), (Somewhat confident), (Very confident))
17. Anything else you would like to add about bots on Twitter? (Free-response)

2.3.2 Labeling with Amazon SageMaker & Mechanical Turk

SageMaker Ground Truth offers the capability to run labeling jobs on three separate types of labeling workforces: private workforce testing, MTurk public workforce (what we used for real job launch), and "vendor" – higher-touch 3rd party services. The generated HTML files do not contain profile photos because not all accounts in the sample still have active profile photos still hosted on Twitter, and we wanted to avoid biasing annotators towards or away from certain accounts due to the presence or absence of a profile photo. Adding profile photo analysis (i.e. by building a CNN) would be a great extension for future work.

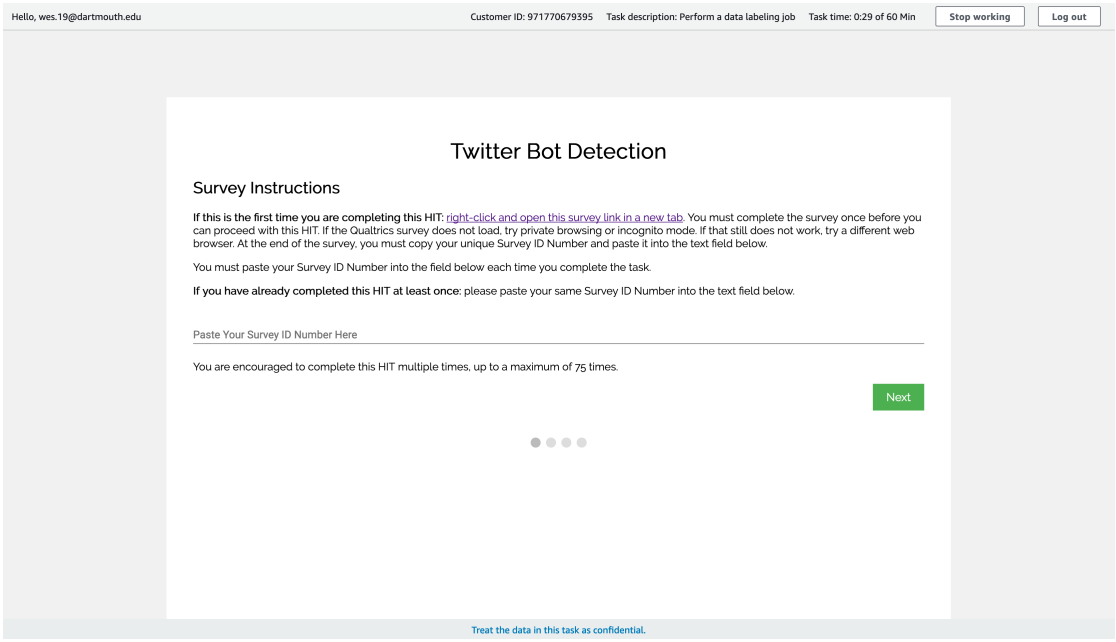


Figure 2.11 Page 1: Insert Survey ID Number after completing the Qualtrics Survey

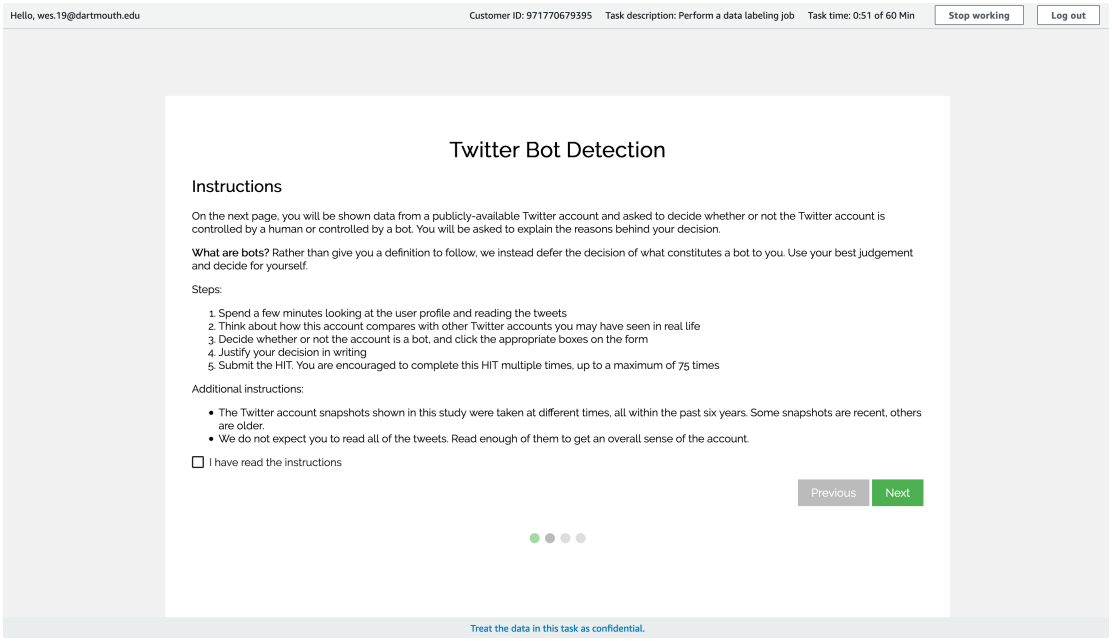


Figure 2.12 Page 2: Task Instructions

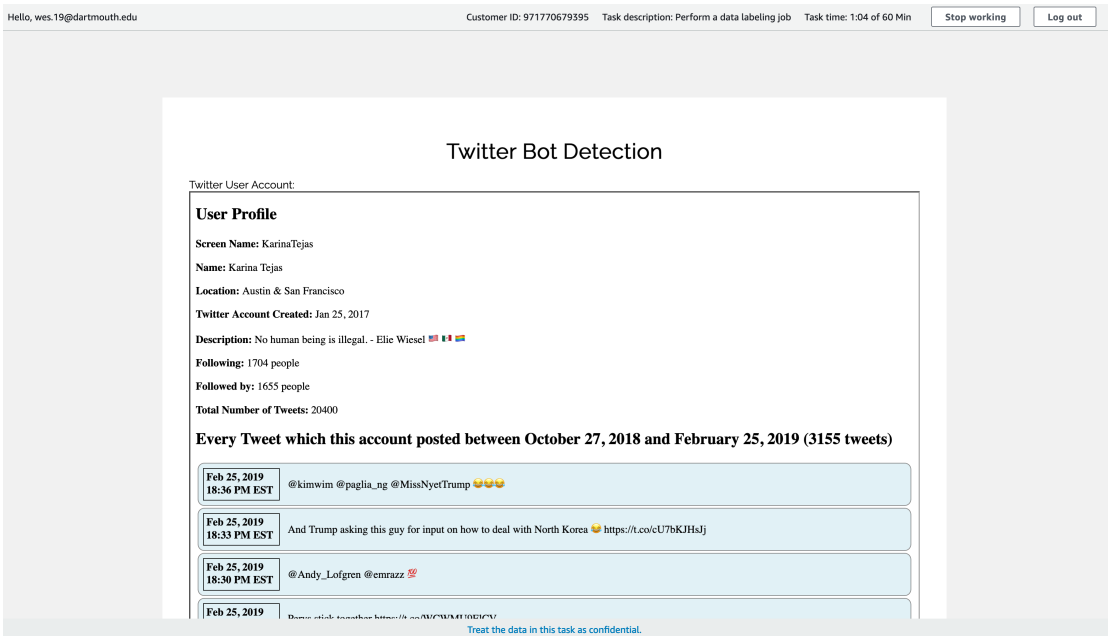


Figure 2.13 Page 3: Top of main page. Notice the iframe, which displays the Twitter account as a nested HTML file

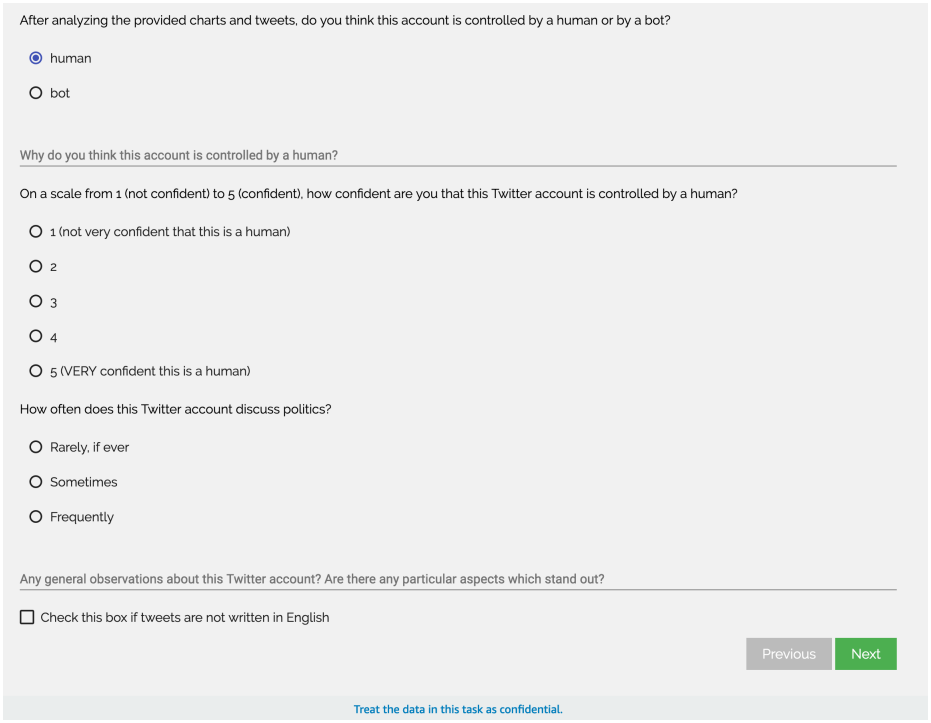


Figure 2.14 Page 3: Conditional form display – "Human" Selected

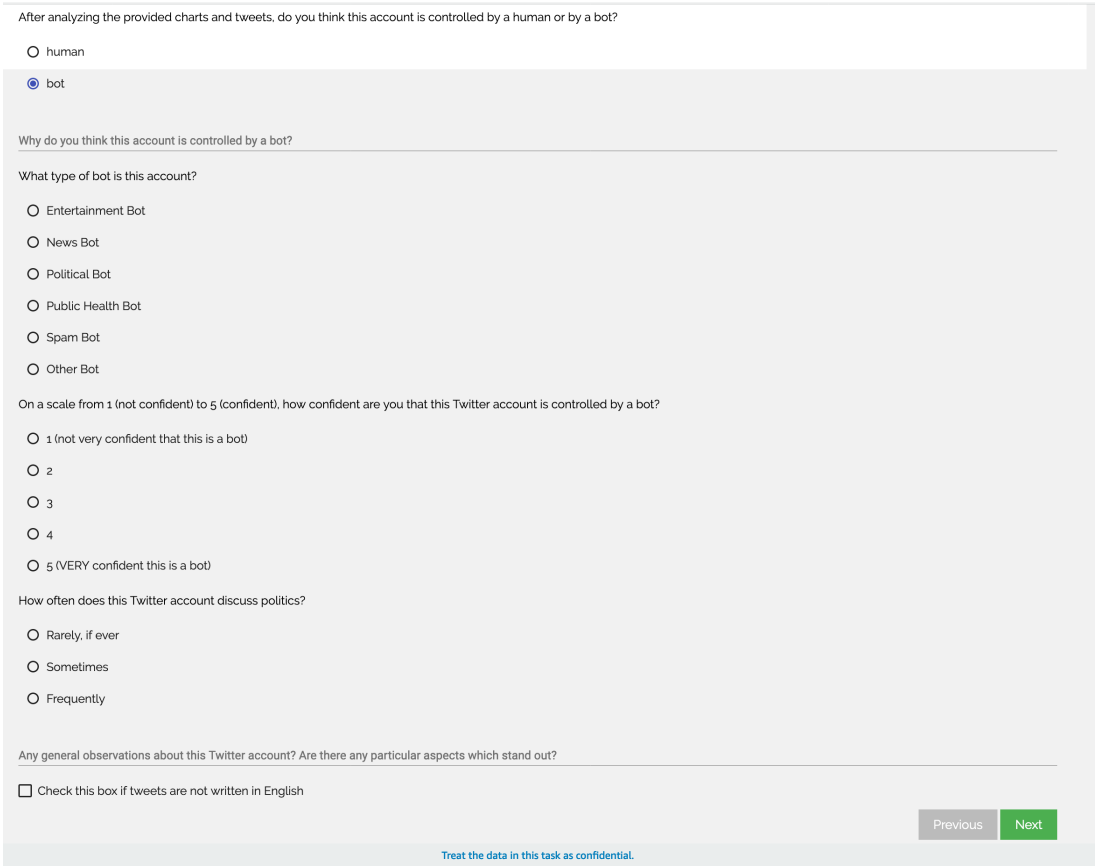


Figure 2.15 Page 3: Conditional form display – "Bot" Selected. Displays list of available bot types

Bot subtypes:

- Entertainment Bot
- News Bot
- Political Bot
- Public Health Bot
  - (the bot type category that all of the BCD 2015 accounts would ideally be classified under.)
- Spam Bot
- Other Bot

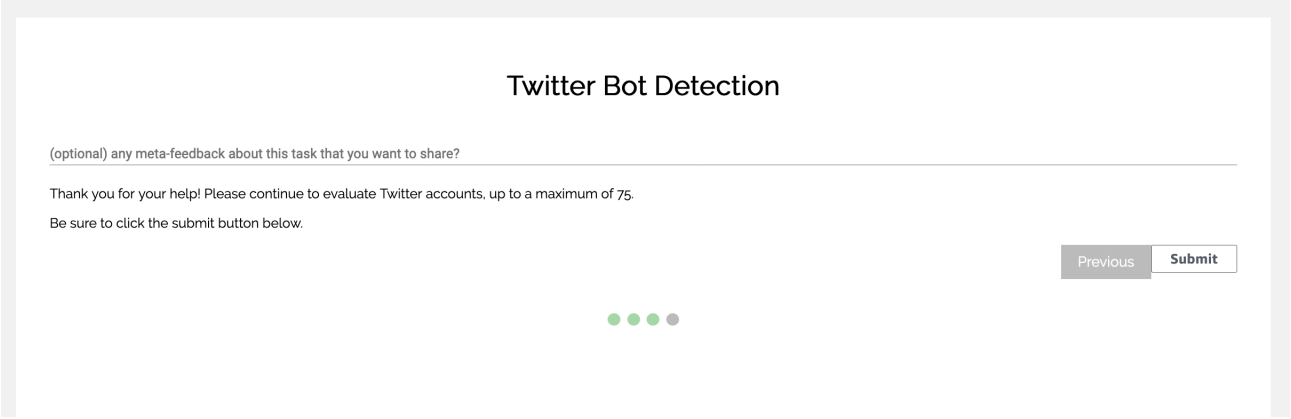


Figure 2.16 Page 4: optional meta-feedback field (Refer to Appendix B for the list of responses to this question)

2.3.3 Study Launch

We were provided a study budget, and our objective was to maximize the number of accounts analyzed in the study without going over budget. We decided to run 3 annotators per account to get a sense of inter-rater reliability, and we decided to pay annotators $.036/account for 90 seconds of their time, quantity of time chosen based on user testing with peers. We knew we would include all 39 bot accounts

from BCD 2015, and we included 117 additional not-known-bot accounts from BCD 2015 for comparison. Our budget thus allowed us to include 1500 additional accounts from the Senate retrospective dataset. To that effect, we scanned through the randomly-selected accounts in our database until we generated 1500 accounts as follows: a total of 1599 accounts were selected randomly, of which 1500 senate accounts (93.87%) passed the filter and were analyzed by MTurkers. One empty (deleted) account was manually included as an experiment, overriding the filter, as an experiment in seeing how annotators would respond to a blank account.
The 98 accounts which did not pass the thresholds were filtered out according to the following methodology:

- 41 blank accounts (2.56%) (either deleted or made private)
- 57 non-English accounts (3.56%) (defined as where <20% of tweets in an account with more than 10 tweets were in English)

Thus, the study consisted of 1,657 total accounts:

- 1501 accounts out of the 85,000 Senate Dataset accounts
- 156 accounts from BCD 2015 (25% known bots, 75% unknown type)

3 annotators per account, paid $0.36/account for a 90 second task

Chapter 3

Results

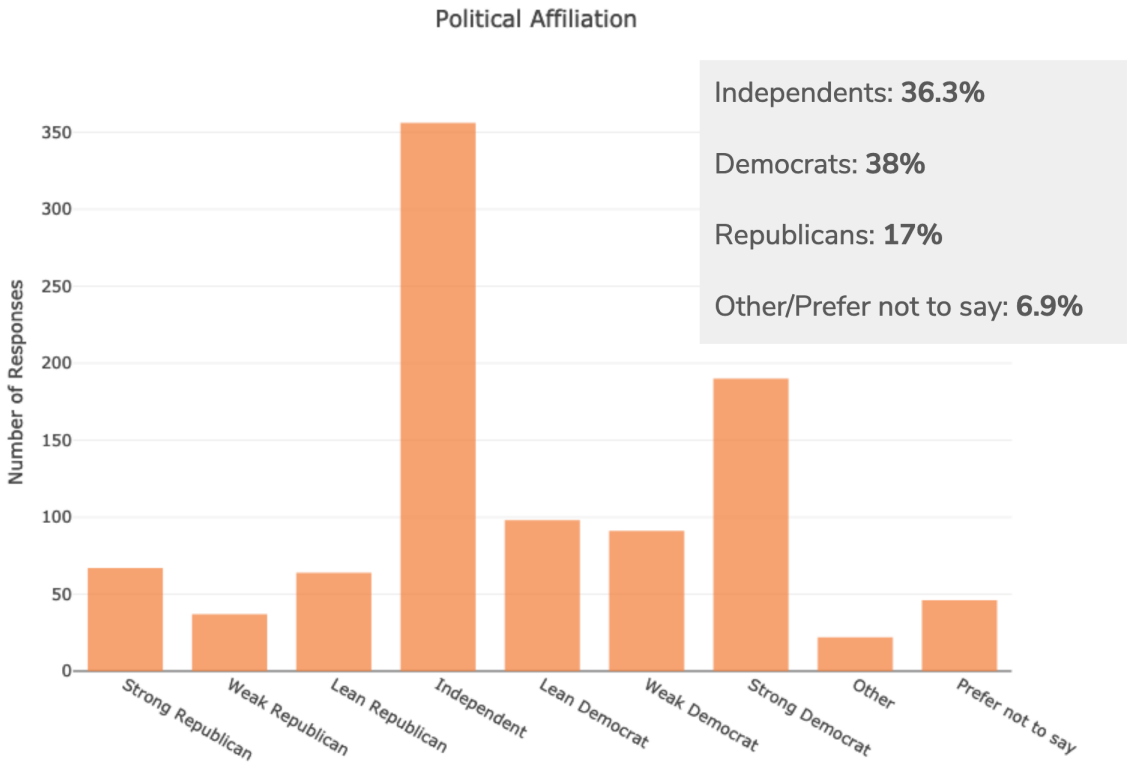
We launched the SageMaker+Qualtrics study on Sunday, May 12th, and the study completed in less than 15 hours. Analysis of the results follows below.

3.1 Qualtrics Demographic Survey Results

- 980 survey respondents
- Mean age: 34.8
- 55% Female, 41% Male

Word	Frequency	Word	Frequency	Word	Frequency
automated	143	program	138	fake	138
computer	115	software	115	automatically	79
information	75	spam	48	robot	40

Figure 3.1 Most frequent words in response to the Qualtrics free-text question: "How would you define what a bot on Twitter is?"



Note that MTurk workers are not representative of the population at large, they just represent a convenient launch point for a study such as this.

We asked individuals in the Qualtrics survey whether or not they believe bots to be helpful or dangerous, and why, and the results to that question are below:

- 72.5% of respondents believe bots to be mostly dangerous. Below are the top words in the bots-are-dangerous free-response:

Word	Frequency	Word	Frequency	Word	Frequency
spread	202	information	181	false	130
news	100	fake	94	misinformation	84
opinion	84	influence	57	spam	45)

- 27.5% of respondents believe bots to be mostly helpful. Below are the top words in the bots-are-helpful free response:

Word	Frequency	Word	Frequency	Word	Frequency
information	39	help	35	news	30
useful	16	without	11	software	11
provide	10	automatically	10	content	10

3.2 MTurk Annotation Responses

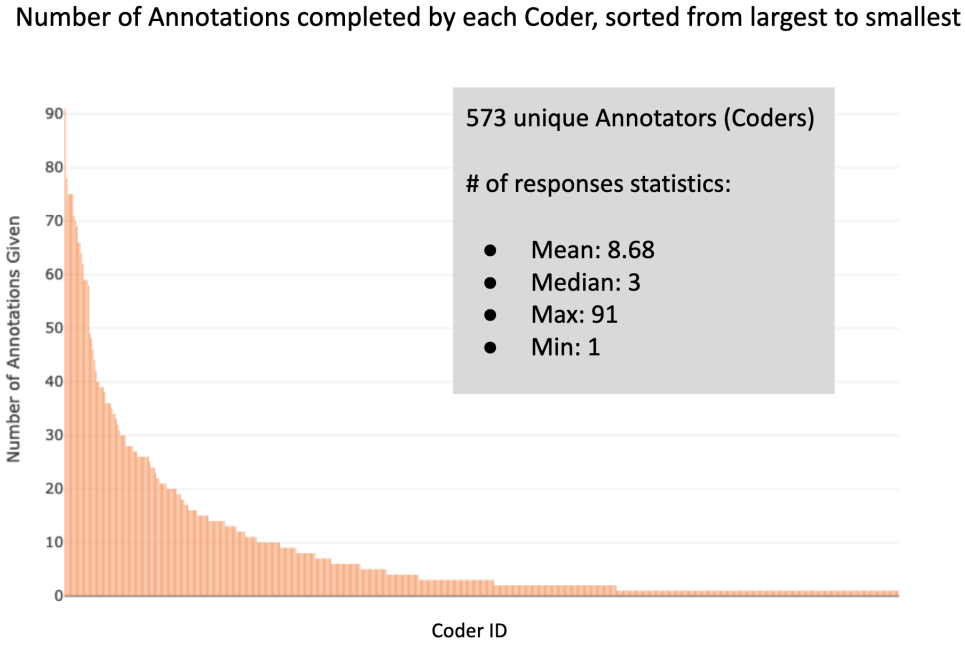


Figure 3.2 Significant power-law drop-off in the number of responses that annotators completed.

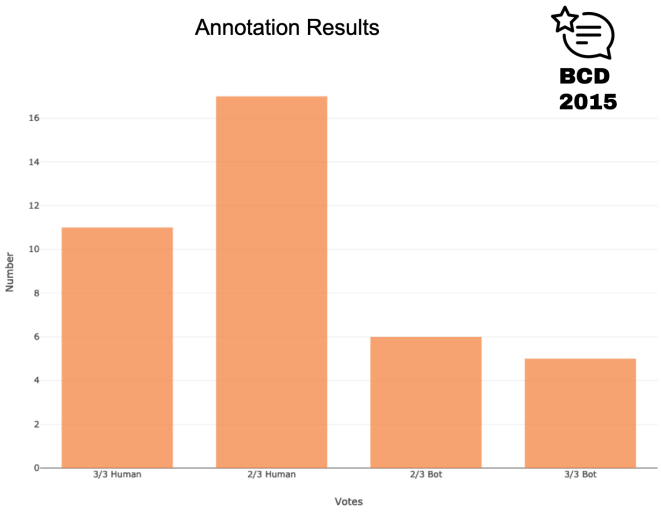
3.3 Ground Truth Results on 39 known bots in BCD 2015

Out of the 39 known BCD 2015 bots:

12.8% (5 out of 39) were detected by all 3 annotators

28.2% (11 out of 39) were detected by at least 2 out of 3 annotators

28.2% (11 out of 39) received unanimous "3/3 Human" votes and escaped detection entirely



Bot detection results on the known ground-truth BCD 2015 dataset

3.3.1 Example BCD 2015 Account

Annotator #1:
"The tweets are weird. especially thanks for reaching out!"
Bot confidence score: 4 / 5
Bot type: Spam Bot

Annotator #2:
"They use the same responses over and over again to respond to many different people."
Bot confidence score: 4 / 5
Bot type: Other Bot

Annotator #3:
"Tweets do not appear following a pattern, talks about different subjects, does not seem to spam anything in tweets."
Human confidence score: 4 / 5

Label: Bot
Confidence Score: 4 / 15

User Profile

Screen Name: HenryBastion
Name: Henry Bastion
Location:
Description:
Following: 57 people
Followed by: 4 people
Total Number of Tweets: 1117

Every Tweet which this account posted between February 25, 2015 and March 12, 2015 (1117 tweets)

Mar 12, 2015 16:43 PM EST	@QRMilitia Radically glorious!
Mar 12, 2015 16:42 PM EST	@DirtyUnkuls RT @aqv21: #AlSharpton Owes the #IRS $4.5 Million, but is not
Mar 12, 2015 16:35 PM EST	RT @_webkids: "@vourneengavin: http://t.co/Ws1Vitkbue" #parents #children #SEN #autism #educateyourself
Mar 12, 2015 16:34 PM EST	RT @LitMargaretNan: Good night y'all! Much love.
Mar 12, 2015 16:33 PM EST	RT @lbakk_money: recipe: "Take a human desire, preferably one that has been around for a really long time and use modern technology to take…
Mar 12, 2015 16:31 PM EST	@CallidoraBeach Thanks for that.
Mar 12, 2015 16:31 PM EST	@remembercsdl Haha

BCD 2015

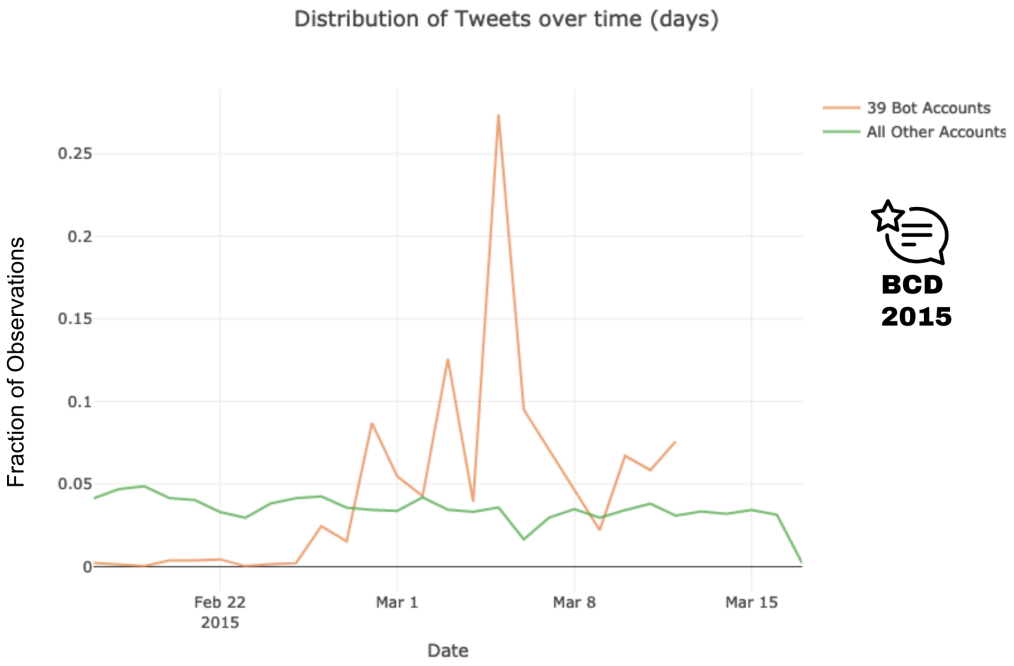


Figure 3.3 BCD Dataset: Tweets over time

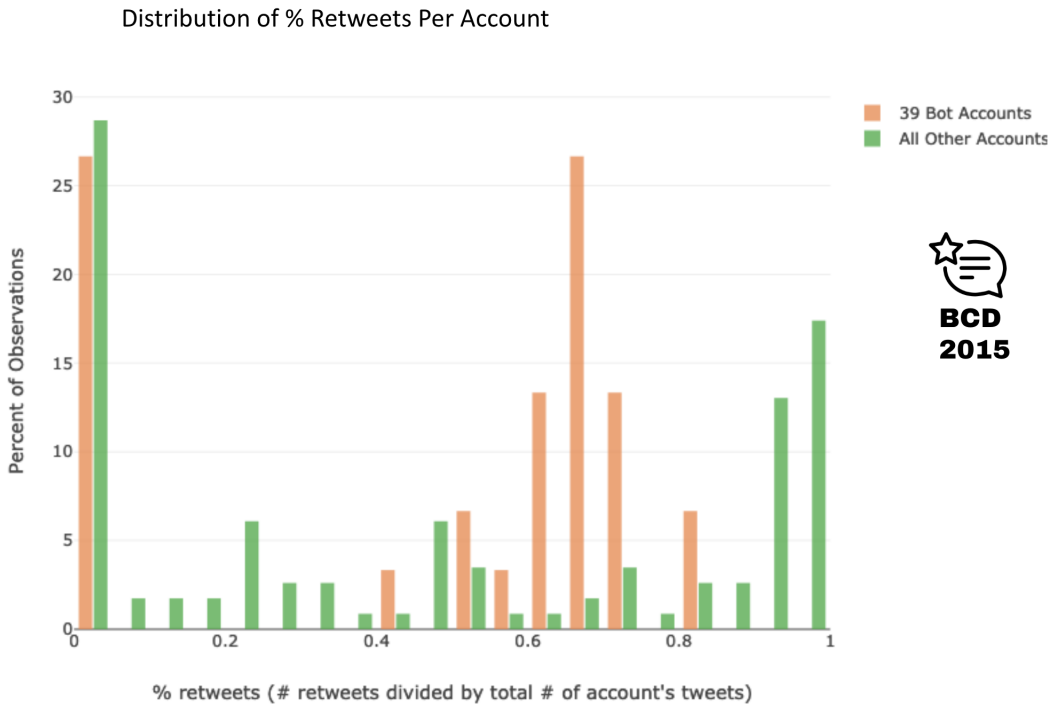


Figure 3.4 BCD Dataset: Retweet distribution. Notice how many of the bot accounts judiciously position themselves in the middle of the distribution, avoiding the high-end of the chart, in contrast with what MTurk labelers believe bots look like.

3.4 Senate Dataset Bot Detection

We write analysis code in Python which allows us to pass in any two lists of Twitter account ids and generate summary statistics and dual-trace charts as output. We develop a negative-15-to-positive-15 confidence score plot. Negative 15 means all three annotators each clicked on "Human" with a confidence of 5, thus (-1)*3*5 = -15. Positive 15 means that all three annotators each clicked on "Bot" with confidence scores of 5 each, hence (1)*3*5 = 15. We plotted the distributions, refer to the following charts. Then, we pick two arbitrary threshold levels to split the two classes at:

1. A threshold set at greater than or equal to "2/3 Bot" vote gives us 358 believed-to-be-bot accounts, or 23.87% of the 1500 senate-account sample.

2. An even higher standard, an 11/15 confidence gives us 87 believed-to-be-bot accounts , or 5.8% of the 1500 senate-account sample.

Recall that these percentages are based on our dataset, which is skewed towards political accounts. These percentages do not generalize beyond this sample. The following charts visualize a split at the 11/15 bot confidence threshold, but many other threshold lines are of course also possible.

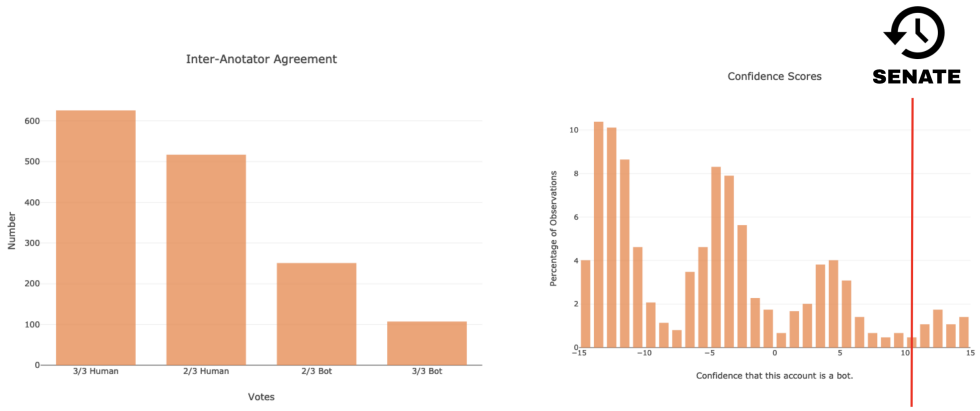


Figure 3.5 Plotting the arbitrary choice of a 11/15 bot/not-bot split

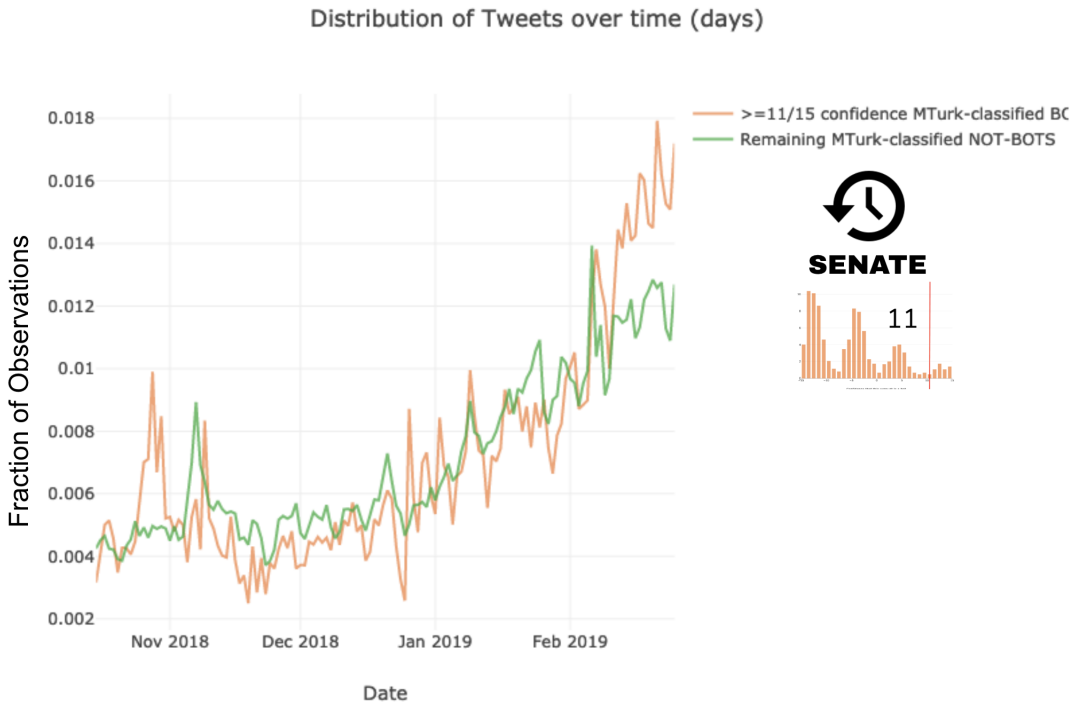


Figure 3.6 Accounts from the Senate retrospective dataset which tweet more frequently are more likely to be perceived by MTurk annotators to be bots. We don't know if this is a correct judgement, only that it is a judgement they are making.

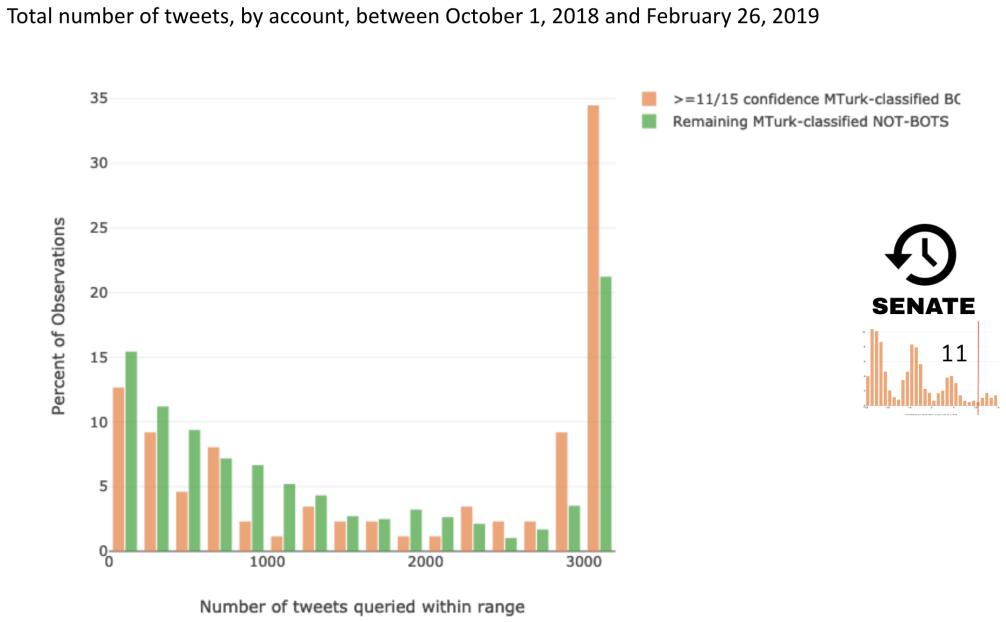


Figure 3.7 Accounts that tweet more frequently are more likely to be perceived to be bots by MTurkers. Note that due to API limitations almost 35% of accounts have so many tweets that we were unable to collect data on them going back all the way to October, hence why so many cap out around 3,000 tweets.

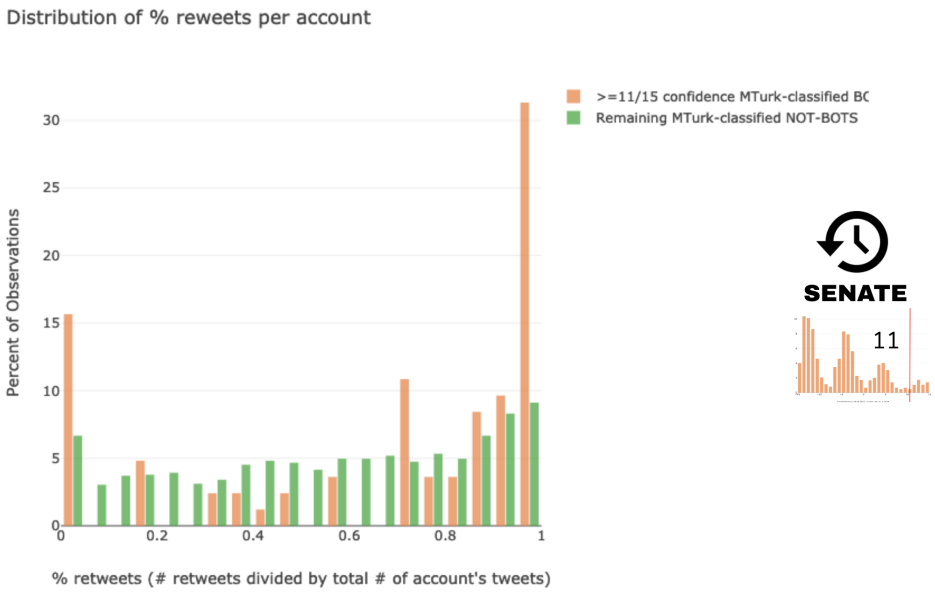


Figure 3.8 Accounts which retweet more frequently are more likely to be perceived to be bots by the MTurk annotators. Also notable: the spike of accounts which never retweet or have zero tweets (chart includes the accounts which were deleted or made private)

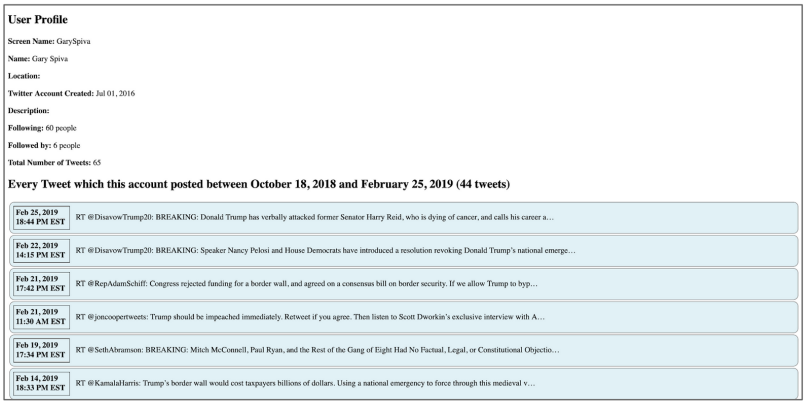
3.4.1 Example Account from Senate Dataset

Annotator #1:
"It's mostly retweets, because it's hard for a bot to post original content without it being obvious it's a bot."
Bot confidence score: 4 / 5
Bot type: News Bot

Annotator #2:
"Many retweets, not a lot of personal comments."
Bot confidence score: 3 / 5
Bot type: Political Bot

Annotator #3:
"Spamming talking about one specific thing in all tweets looks fake"
Bot confidence score: 4 / 5
Bot type: News Bot

Label: Bot
Confidence Score: 11 / 15



User Profile

Screen Name: GarySpiva
Name: Gary Spiva
Location:
Twitter Account Created: Jul 01, 2016
Description:
Following: 60 people
Followed by: 6 people
Total Number of Tweets: 65

Every Tweet which this account posted between October 18, 2018 and February 25, 2019 (44 tweets)

Feb 25, 2019 18:44 PM EST	RT @DisavowTrump20: BREAKING: Donald Trump has verbally attacked former Senator Harry Reid, who is dying of cancer, and calls his career a…
Feb 22, 2019 14:15 PM EST	RT @DisavowTrump20: BREAKING: Speaker Nancy Pelosi and House Democrats have introduced a resolution revoking Donald Trump's national emerge…
Feb 21, 2019 17:42 PM EST	RT @RepAdamSchiff: Congress rejected funding for a border wall, and agreed on a consensus bill on border security. If we allow Trump to byp…
Feb 21, 2019 11:30 AM EST	RT @joncoopertweets: Trump should be impeached immediately. Retweet if you agree. Then listen to Scott Dworkin's exclusive interview with A…
Feb 19, 2019 17:34 PM EST	RT @SethAbramson: BREAKING: Mitch McConnell, Paul Ryan, and the Rest of the Gang of Eight Had No Factual, Legal, or Constitutional Objectio…
Feb 14, 2019 18:33 PM EST	RT @KamalaHarris: Trump's border wall would cost taxpayers billions of dollars. Using a national emergency to force through this medieval v…

SENATE

To be very clear – we do not know if this is a bot. It may just be a partisan who retweets frequently. (@GarySpiva, if you're reading this, reach out and let us know.)

Chapter 4

Discussion

A few remarks regarding future work: with only 39 known ground-truth bots in the dataset, it would be interesting to have more people look at the same accounts, or incorporate other known ground-truth datasets. We would not want a hypothetical bot detector system in production falsely accusing accounts of being bots.
Possible improvements to MTurk Labeling Task going forward include:

1. Incorporate profile images, inline media, conversations between accounts
2. Add graph network features (although Twitter rate limits make this hard)
3. Pay for more than just 90 seconds per Twitter account
4. Increase the number of coders per account (5, 7, or 9 instead of just 3)

Improvements to the Data Archive Service:

1. account-level real-time collection
2. build a real-time classifier

More in-depth analysis:

1. Control for:
    (a) annotator bias

(b) number of tweets

(c) political affiliation

Going forward, it will be interesting to investigate the effects of partisanship, as well as demographic characteristics on how they affect human tendency to judge accounts as bots or humans. Furthermore, it would be interesting to apply more advanced NLP – bag-of-words analysis only scratches the surface. As a next step, incorporating the "bot subtype" data into a multi-class classifier would be an interseting extension to this work. This dataset and this research opens doors for more collaboration across both Computer Science and Political Science.

Future work:

- Improve collection methodology (pull full profiles in real-time, larger sample, more search keywords)
- Improve MTurk survey interface (add profile images, add inline media, add conversation context)
- Run an "analyst-style" job where we give MTurkers a much more robust set of tools and features to look at (eg derived plots and graphs that they wouldn't otherwise see while browsing Twitter)
- Build a "what people think are bots" classifier
- Build a bot classifier
- Rerun identical task months or years later, to see if/how responses change over time, see whether or not accounts are labeled again as bots, see if accounts have been blocked or deleted themselves.
- Look at accounts again, see how many remain active and have not been blocked

Chapter 5

Conclusions

Bots are constantly changing. As soon as an academic builds an accurate bot detector, botmakers can build bots designed to get around the new detectors, in a never-ending sequence of cat versus mouse.

It is rarely clear what is human and what isn't, especially to an untrained observer. The ground truth results using the BCD 2015 dataset representing a humbling reminder of how hard it can be to detect bots. Roughly a quarter were detected (11/39); just as many flew under the radar of all three annotators.

When it comes to the Senate accounts, we observed earlier that the annotators in this study were more likely to believe that a high-frequency posting, high percentage of retweets out of total number of tweets account is a bot. While this may be true, there is a risk that annotators might just be demonizing the highest-visibility, highest-frequency-posting people, and falsely caricaturing them as bots, when they might actually just be partisans with different political opinions. Going forward, we need to do additional NLP analysis of the actual tweet content in order to get a better understanding of these accounts.

It is a small sample, and MTurk populations are known to be skewed, but nevertheless it is an interesting snapshot into one particular batch of ground-truth bots.

In summary: we learned that certain political candidates are discussed on Twitter much more frequently than others. We learned that humans were not very good at detecting these particular BCD 2015 bots. As a possible follow-up, we could provide

the next iteration of annotators with more metadata and powerful analyst-style charts to see what they could come up with when provided interactive visualizations of different account data and metadata.

It is likely that we will discover additional attributes as we continue our analysis of this dataset going forward. We want to avoid unintentionally building a detector which simply flags highly-political accounts. Humans, not just bots, also express polar (as well as moderate) political opinions, and it is important that we do not inadvertently limit their freedom of speech by mischaracterizing them as bots.

We will conclude with two closing observations. First, we note that data collection is not an implementation detail. Finally, we note that a model is only as good as its underlying ground truth, and for bots out in the wild, ground truth is in short supply.

References

[1] C. A. Davis, O. Varol, E. Ferrara, A. Flammini, and F. Menczer, "Botornot: A system to evaluate social bots", in *Proceedings of the 25th International Conference Companion on World Wide Web*, International World Wide Web Conferences Steering Committee, 2016, pp. 273–274.

[2] K.-C. Yang, O. Varol, C. A. Davis, E. Ferrara, A. Flammini, and F. Menczer, "Arming the public with artificial intelligence to counter social bots", *Human Behavior and Emerging Technologies*, vol. 1, no. 1, pp. 48–61, 2019.

[3] V. Subrahmanian, A. Azaria, S. Durst, V. Kagan, A. Galstyan, K. Lerman, L. Zhu, E. Ferrara, A. Flammini, and F. Menczer, "The DARPA Twitter bot challenge", *Computer*, vol. 49, no. 6, pp. 38–46, 2016.

[4] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, "The rise of social bots", *Communications of the ACM*, vol. 59, no. 7, pp. 96–104, 2016.

[5] T. Finin, W. Murnane, A. Karandikar, N. Keller, J. Martineau, and M. Dredze, "Annotating named entities in twitter data with crowdsourcing", in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, Association for Computational Linguistics, 2010, pp. 80–88.

[6] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia, "Who is tweeting on twitter: Human, bot, or cyborg?", in *Proceedings of the 26th annual computer security applications conference*, ACM, 2010, pp. 21–30.

[7] R. Cohen and D. Ruths, "Classifying political orientation on Twitter: It's not easy!", in *Seventh International AAAI Conference on Weblogs and Social Media*, 2013.

[8] *MIT Election Data Science Lab*. [Online]. Available: `https://electionlab.mit.edu/data`.

[9] M. Honnibal and I. Montani, *Prodigy: Radically efficient machine teaching. An annotation tool powered by active learning.* [Online]. Available: `https://prodi.gy/` (visited on May 20, 2019).

[10] *Amazon SageMaker Ground Truth*. [Online]. Available: `https://aws.amazon.com/sagemaker/groundtruth/`.

[11] *MTurk introduces Crowd HTML Elements, a library of easy-to-use task interfaces for bounding box, semantic segmentation, classification and more*, January 2019. [Online]. Available: `https://blog.mturk.com/mturk-introduces-crowd-html-elements-a-library-of-easy-to-use-task-interfaces-for-bounding-box-35bb9c860069`.

[12] J. J. Dressel, "Accuracy and Racial Biases of Recidivism Prediction Instruments", Dartmouth College, Computer Science, Hanover, NH, Tech. Rep. TR2017-822, Jun. 2017. [Online]. Available: `http://www.cs.dartmouth.edu/reports/TR2017-822.pdf`.

[13] *More about restricted uses of the Twitter APIs.* [Online]. Available: `https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases.html`.

[14] M. R. Lawson, "The Next Generation of EMPRESS: A Metadata Management System For Accelerated Scientific Discovery at Exascale", Dartmouth College, Computer Science, Hanover, NH, Tech. Rep. TR2018-846, Jun. 2018. [Online]. Available: `http://www.cs.dartmouth.edu/%20trdata/reports/TR2018-846.pdf`.

[15] *How TO - Form with Multiple Steps.* [Online]. Available: `https://www.w3schools.com/howto/howto_js_form_steps.asp`.

[16] L. Marujo, W. Ling, I. Trancoso, C. Dyer, A. W. Black, A. Gershman, D. M. de Matos, J. Neto, and J. Carbonell, "Automatic keyword extraction on twitter", in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, vol. 2, 2015, pp. 637–643.

[17] *TWARC: A command line tool (and Python library) for archiving Twitter JSON.* [Online]. Available: `https://github.com/DocNow/twarc` (visited on May 20, 2019).

[18] J. Roesslein, *Tweepy: An easy-to-use Python library for accessing the Twitter API.* https://github.com/tweepy/tweepy. [Online]. Available: `https://www.tweepy.org/`.

[19] E. Lancaster, T. Chakraborty, and V. Subrahmanian, "Maltp: Parallel prediction of malicious tweets", *IEEE Transactions on Computational Social Systems*, vol. 5, no. 4, pp. 1096–1108, 2018.

[20] P. Vijayaraghavan, S. Vosoughi, and D. Roy, "Automatic detection and categorization of election-related tweets", in *Tenth International AAAI Conference on Web and Social Media*, 2016.

[21] N. Gillani, A. Yuan, M. Saveski, S. Vosoughi, and D. Roy, "Me, my echo chamber, and I: Introspection on social media polarization", in *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, International World Wide Web Conferences Steering Committee, 2018, pp. 823–831.

[22] M. D. Conover, B. Gonçalves, J. Ratkiewicz, A. Flammini, and F. Menczer, "Predicting the political alignment of twitter users", in *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*, IEEE, 2011, pp. 192–199.

[23] A. Bakliwal, J. Foster, J. van der Puil, R. O'Brien, L. Tounsi, and M. Hughes, "Sentiment analysis of political tweets: Towards an accurate classifier", Association for Computational Linguistics, 2013.

Appendix A

Twitter API Search Keywords

List of Twitter API search terms used to construct the real-time Senate database. The @ username search only downloads tweets posted by the candidates themselves, unlike the other two searches. 'inc' stands for 'incumbent'.

Arizona Senate Race **Kyrsten Sinema (D) versus Angela Green (G) versus Martha McSally (R)**	
Search Term	*Search Type*
@kyrstensinema	Username
@RepSinema	Username
@RepMcSally	Username
@MarthaMcSally	Username
#KyrstenSinema	Hashtag
Kyrsten Sinema	String
#AngelaGreen	Hashtag
Angela Green	String
#MarthaMcSally	Hashtag
Martha McSally	String

California Senate Race **Dianne Feinstein (D inc) versus Kevin de Len (D)**	
Search Term	*Search Type*
@SenFeinstein	Username
@DianneFeinstein	Username
@kdeleon	Username
#DianneFeinstein	Hashtag
Dianne Feinstein	String
#KevinDeLen	Hashtag
Kevin de Len	String

Connecticut Senate Race **Chris Murphy (D inc) versus Matthew Corey (R)**	
Search Term	*Search Type*
@ChrisMurphyCT	Username
@MattCoreyCT	Username
#ChrisMurphy	Hashtag
Chris Murphy	String
#MatthewCorey	Hashtag
Matthew Corey	String

Delaware Senate Race **Tom Carper (D inc) versus Rob Arlett (R) versus Nadine Frost (L)**	
Search Term	*Search Type*
@SenatorCarper	Username
@TomCarperforDE	Username
@RobArlett	Username
#Tom Carper	Hashtag
TomCarper	String
#RobArlett	Hashtag
Rob Arlett	String
#NadineFrost	Hashtag
Nadine Frost	String

Florida Senate Race **Bill Nelson (D) versus Rick Scott (R)**	
Search Term	*Search Type*
@SenBillNelson	Username
@NelsonForSenate	Username
@FLGovScott	Username
@ScottforFlorida	Username
#BillNelson	Hashtag
Bill Nelson	String
#RickScott	Hashtag
Rick Scott	String

Hawaii Senate Race **Mazie Hirono (D inc) Ron Curtis (R)**	
Search Term	*Search Type*
@mazieforhawaii	Username
@maziehirono	Username
@rcurtis808	Username
#MazieHirono	Hashtag
Mazie Hirono	String
#RonCurtis	Hashtag
Ron Curtis	String

Indiana Senate Race **Joe Donnelly (D) versus Mike Braun (R)**	
Search Term	*Search Type*
@SenDonnelly	Username
@JoeforIndiana	Username
@braun4indiana	Username
#JoeDonnelly	Hashtag
Joe Donnelly	String
#MikeBraun	Hashtag
Mike Braun	String

Maine Senate Race **Angus King (I inc) versus Zak Ringelstein (D) versus Eric Brakey (R)**	
Search Term	*Search Type*
@SenAngusKing	Username
@RingelsteinME	Username
@SenatorBrakey	Username
#AngusKing	Hashtag
Angus King	String
#ZakRingelstein	Hashtag
Zak Ringelstein	String
#EricBrakey	Hashtag
Eric Brakey	String

Maryland Senate Race **Ben Cardin (D inc) versus Tony Campbell (R)**	
Search Term	*Search Type*
@BenCardinforMD	Username
@SenatorCardin	Username
@Campbell4MD	Username
#BenCardin	Hashtag
Ben Cardin	String
#TonyCampbell	Hashtag
Tony Campbell	String

Massachusetts Senate Race **Elizabeth Warren (D inc) versus Shiva Ayyadurai (I) versus Geoff Diehl (R)**	
Search Term	*Search Type*
@SenWarren	Username
@elizabethforma	Username
@va_shiva	Username
@RepGeoffDiehl	Username
@DiehlForSenate	Username
#ElizabethWarren	Hashtag
Elizabeth Warren	String
#ShivaAyyadurai	Hashtag
Shiva Ayyadurai	String
#GeoffDiehl	Hashtag
Geoff Diehl	String

Michigan Senate Race **Debbie Stabenow (D inc) versus John James (R)**	
Search Term	*Search Type*
@stabenow	Username
@SenStabenow	Username
@JohnJamesMI	Username
#DebbieStabenow	Hashtag
Debbie Stabenow	String
#JohnJames	Hashtag
John James	String

Minnesota (Special Election) Senate Race **Tina Smith (D inc) versus Karin Housley (R)**	
Search Term	*Search Type*
@TinaSmithMN	Username
@SenTinaSmith	Username
@KarinHousley	Username
#TinaSmith	Hashtag
Tina Smith	String
#KarinHousley	Hashtag
Karin Housley	String

Minnesota Senate Race **Amy Klobuchar (D inc) versus Jim Newberger (R)**	
Search Term	*Search Type*
@amyklobuchar	Username
@SenAmyKlobuchar	Username
@NewbergerJim	Username
#AmyKlobuchar	Hashtag
Amy Klobuchar	String
#JimNewberger	Hashtag
Jim Newberger	String

Mississippi (Special Election) Senate Race **Cindy Hyde-Smith (R, inc) versus Mike Espy (D)**	
Search Term	*Search Type*
@cindyhydesmith	Username
@SenHydeSmith	Username
@espyforsenate	Username
#CindyHyde-Smith	Hashtag
Cindy Hyde-Smith	String
#MikeEspy	Hashtag
Mike Espy	String

Mississippi Senate Race **Roger F. Wicker (R inc) versus David Baria (D)**	
Search Term	*Search Type*
@RogerWicker	Username
@SenatorWicker	Username
@dbaria	Username
#RogerWicker	Hashtag
Roger Wicker	String
#DavidBaria	Hashtag
David Baria	String

Missouri Senate Race **Claire McCaskill versus Josh Hawley**	
Search Term	*Search Type*
@clairecmc	Username
@McCaskill4MO	Username
@HawleyMO	Username
#ClaireMcCaskill	Hashtag
Claire McCaskill	String
#JoshHawley	Hashtag
Josh Hawley	String

Montana Senate Race **Jon Tester (D inc) versus Matt Rosendale (R)**	
Search Term	*Search Type*
@jontester	Username
@SenatorTester	Username
@MattForMontana	Username
@MattRosendale	Username
#JonTester	Hashtag
Jon Tester	String
#MattRosendale	Hashtag
Matt Rosendale	String

Nebraska Senate Race **Deb Fischer (R inc) versus Jane Raybould (D)**	
Search Term	*Search Type*
@SenatorFischer	Username
@DebFischerNE	Username
@JaneRaybould	Username
#DebFischer	Hashtag
Deb Fischer	String
#JaneRaybould	Hashtag
Jane Raybould	String

Nevada Senate Race **Dean Heller (R inc) versus Jacky Rosen (D)**	
Search Term	*Search Type*
@DeanHeller	Username
@SenDeanHeller	Username
@RosenforNevada	Username
@RepJackyRosen	Username
#DeanHeller	Hashtag
Dean Heller	String
#JackyRosen	Hashtag
Jacky Rosen	String

New Jersey Senate Race **Bob Menendez (D inc) versus Bob Hugin (R)**	
Search Term	*Search Type*
@SenatorMenendez	Username
@BobMenendezNJ	Username
@BobHugin	Username
#BobMenendez	Hashtag
Bob Menendez	String
#BobHugin	Hashtag
Bob Hugin	String

New Mexico Senate Race **Martin Heinrich (D inc) versus Gary Johnson (I) versus Mick Rich (R)**	
Search Term	*Search Type*
@TeamHeinrich	Username
@MartinHeinrich	Username
@GovGaryJohnson	Username
@MickRich4Senate	Username
#MartinHeinrich	Hashtag
Martin Heinrich	String
#GaryJohnson	Hashtag
Gary Johnson	String
#MickRich	Hashtag
Mick Rich	String

New York Senate Race **Kirsten Gillibrand (D inc) versus Chele Chiavacci Farley (R)**	
Search Term	*Search Type*
@gillibrandny	Username
@SenGillibrand	Username
@CheleFarley	Username
#KirstenGillibrand	Hashtag
Kirsten Gillibrand	String
#CheleChiavacciFarley	Hashtag
Chele Chiavacci Farley	String

North Dakota Senate Race **Heidi Heitkamp versus Kevin Cramer**	
Search Term	*Search Type*
@SenatorHeitkamp	Username
@KevinCramer	Username
@RepKevinCramer	Username
#HeidiHeitkamp	Hashtag
Heidi Heitkamp	String
#KevinCramer	Hashtag
Kevin Cramer	String

Ohio Senate Race **Sherrod Brown (D inc) versus Jim Renacci (R)**	
Search Term	*Search Type*
@SherrodBrown	Username
@SenSherrodBrown	Username
@RepJimRenacci	Username
@JimRenacci	Username
#SherrodBrown	Hashtag
Sherrod Brown	String
#JimRenacci	Hashtag
Jim Renacci	String

Pennsylvania Senate Race **Bob Casey Jr. (D inc) versus Lou Barletta (R)**	
Search Term	*Search Type*
@Bob_Casey	Username
@louforsenate	Username
@RepLouBarletta	Username
#BobCasey	Hashtag
Bob Casey	String
#LouBarletta	Hashtag
Lou Barletta	String

Rhode Island Senate Race **Sheldon Whitehouse (D inc) versus Robert Flanders (R)**	
Search Term	*Search Type*
@SheldonforRI	Username
@SenWhitehouse	Username
@flanders4senate	Username
#SheldonWhitehouse	Hashtag
Sheldon Whitehouse	String
#RobertFlanders	Hashtag
Robert Flanders	String

Tennessee Senate Race **Phil Bredesen (D) versus Marsha Blackburn (R)**	
Search Term	*Search Type*
@PhilBredesen	Username
@MarshaBlackburn	Username
@VoteMarsha	Username
#PhilBredesen	Hashtag
Phil Bredesen	String
#MarshaBlackburn	Hashtag
Marsha Blackburn	String

Texas Senate Race **Ted Cruz (R inc) versus Beto O'Rourke (D)**	
Search Term	*Search Type*
@TeamTedCruz	Username
@SenTedCruz	Username
@tedcruz	Username
@BetoORourke	Username
@RepBetoORourke	Username
#TedCruz	Hashtag
Ted Cruz	String
#BetoORourke	Hashtag
Beto O'Rourke	String

Utah Senate Race **Mitt Romney (R) versus Jenny Wilson (D)**	
Search Term	*Search Type*
@MittRomney	Username
@JennyWilsonUT	Username
#MittRomney	Hashtag
Mitt Romney	String
#JennyWilson	Hashtag
Jenny Wilson	String

Vermont Senate Race **Bernie Sanders (I inc) versus Lawrence Zupan (R)**	
Search Term	*Search Type*
@SenSanders	Username
@BernieSanders	Username
#BernieSanders	Hashtag
Bernie Sanders	String

Virginia Senate Race **Tim Kaine (D inc) versus Corey Stewart (R)**	
Search Term	*Search Type*
@timkaine	Username
@CoreyStewartVA	Username
#TimKaine	Hashtag
Tim Kaine	String
#CoreyStewart	Hashtag
Corey Stewart	String

Washington Senate Race **Maria Cantwell (D inc) versus Susan Hutchison (R)**	
Search Term	*Search Type*
@SenatorCantwell	Username
@MariaCantwell	Username
@Susan4Senate	Username
@Susan_Hutch	Username
#MariaCantwell	Hashtag
Maria Cantwell	String
#SusanHutchison	Hashtag
Susan Hutchison	String

West Virginia Senate Race **Joe Manchin (D inc) versus Patrick Morrisey (R)**	
Search Term	*Search Type*
@Sen$joeManchin$	Username
@JoeManchinWV	Username
@MorriseyWV	Username
#JoeManchin	Hashtag
Joe Manchin	String
#PatrickMorrisey	Hashtag
Patrick Morrisey	String

Wisconsin Senate Race **Tammy Baldwin (D inc) versus Leah Vukmir (R)**	
Search Term	*Search Type*
@tammybaldwin	Username
@SenatorBaldwin	Username
@LeahVukmir	Username
#TammyBaldwin	Hashtag
Tammy Baldwin	String
#LeahVukmir	Hashtag
Leah Vukmir	String

Wyoming Senate Race **John Barrasso (R inc) versus Gary Trauner (D)**	
Search Term	*Search Type*
@barrassoforwyo	Username
@SenJohnBarrasso	Username
@TraunerforWY	Username
#JohnBarrasso	Hashtag
John Barrasso	String
#GaryTrauner	Hashtag
Gary Trauner	String

Appendix B

Mechanical Turk Participant Feedback on Labeling Task

Free-text responses from the survey's optional Meta-feedback field. Responses such as 'no' and 'N/A' were filtered out, but all substantive responses are shown below.

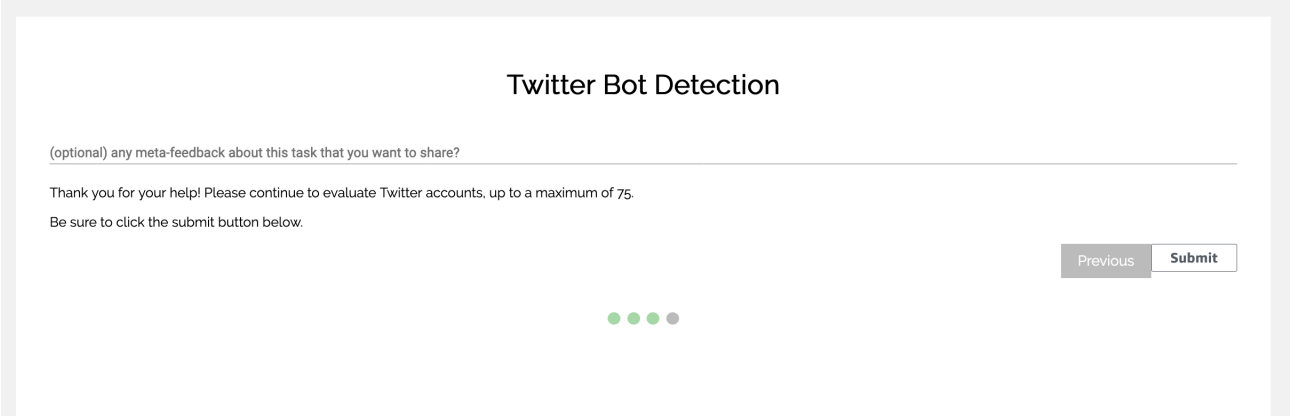


Figure B.1 Meta-feedback field

- no really no so much on politics
- Interesting task, but it will be somewhat limited. Could be a good filter to try to weed out the most obvious bot acc.
- I'm going to stop here because I'm honestly not sure if I'm any good at this. No wonder bots are such a problem, it's really hard to tell.
- good expreinces

- Thank you for the opportunity to participate.
- Having an actual screenshot of the account could help identifying if it's a bot or not.
- add some pictures will be fine
- very interestinge
- I am not seeing the 90 second timer. I'm not sure where it should be at. I hope that I'm doing this right.
- The program didn't catch a private account–no tweets to review.
- Need a button for private and deleted accounts.
- This might be a human, but if so, I think it's a human that is being paid to post several times a day.
- Pay's a bit low considering the time you need to spend to answer this thoughfully.
- This was an interesting task. Thank you!
- good hit.
- I am not seeing the 90 second timer, I feel either it isn't there (which would be a good thing) or fear I might get rejected hits because of this. I don't want to speed and rush my work , as I like to give the best work I can give. I do see a 60 minute timer, just wondering since it was in the instructions.
- Very Nice Survey.
- sisgnificant amount of infor to process in a subjective task
- good hit.
- There could be some guidelines to explain how to identify a bot

- Can not evaluate because account has been deleted or is private
- thank you
- Twitter account was made private so I was unable to make a decision based on account's tweets.
- Nope. Keep up the good work!
- Great task
- no not at this time
- Thank you for this hit
- There is no 90 second timer as mentioned in the instructions. Would also be helpful to see examples of bot accounts, and distinguish between bots and spammers.
- I could not determine whether or not this user was a human or bot because the account is suspended.
- the general observation box is usually redundant.
- I think I just did this one, unless I forgot to submit it the first time?
- again, the "general observation" box is kind of a repetition
- Nope, looks good. Thanks!
- you should add an option for suspended accounts.
- Fun survey!
- add images if anyone is there.
- The tweets were protected. I made my best guess based on the meta data available.

- This was simple and easy to figure out of to complete.
- Might be good to try and sort by language.
- yes', 'This account was apparently deleted, only a user name was left there.
- good survey
- Hardest one I've done yet because there were s little tweets to go off of and it really seems like a bot but makes a few human-like statements. It was hard to choose between bot and human.
- good survey
- Need to click at least one option to continue, no chat script.
- Great task, thank you.
- Thank you for this hit
- The tweets were protected. Unsure if they were an actual bot or not.
- Pro-democratic person
- It did not show their tweets on the page before. But I looked them up on twitter and it showed them there.'...